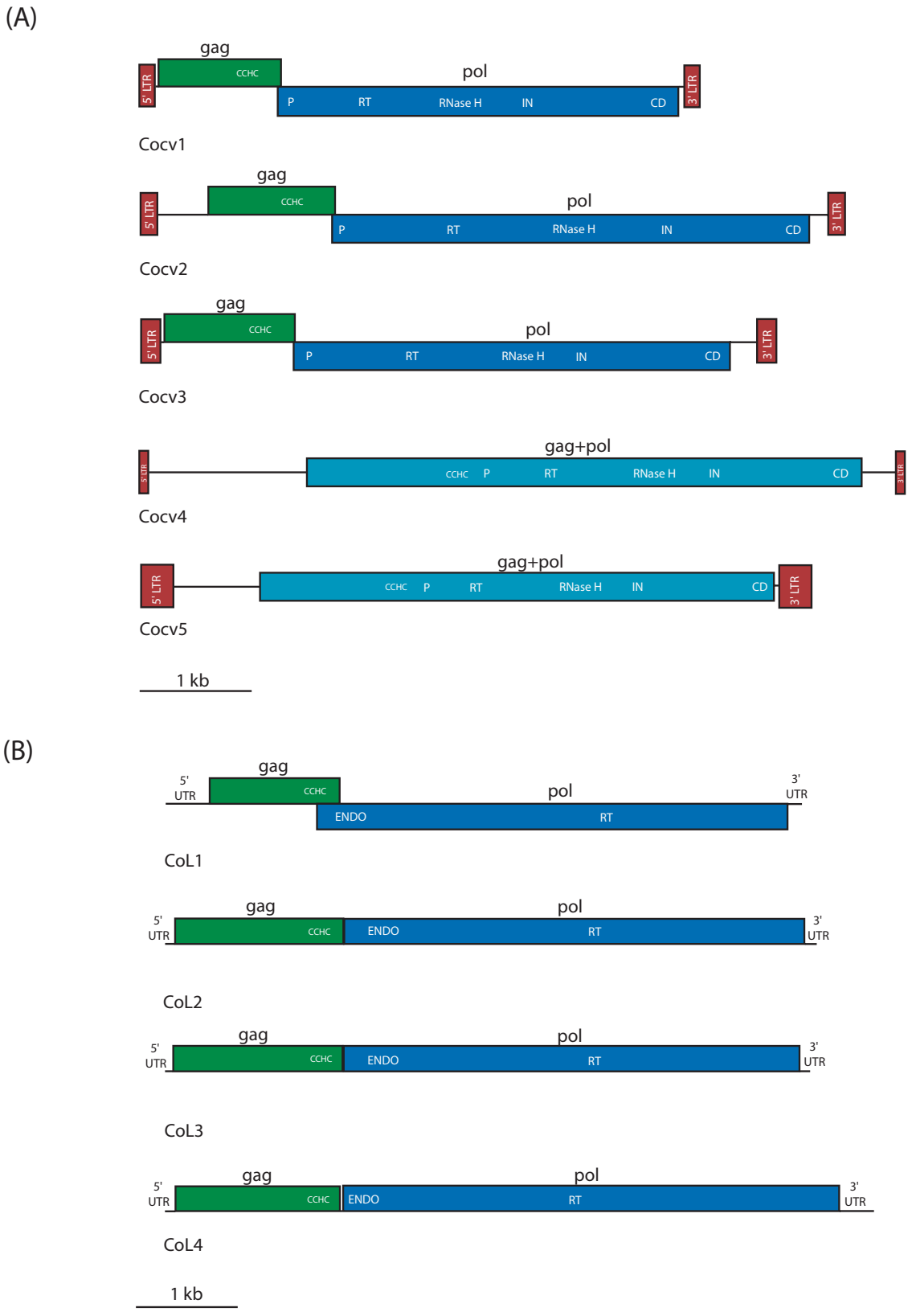
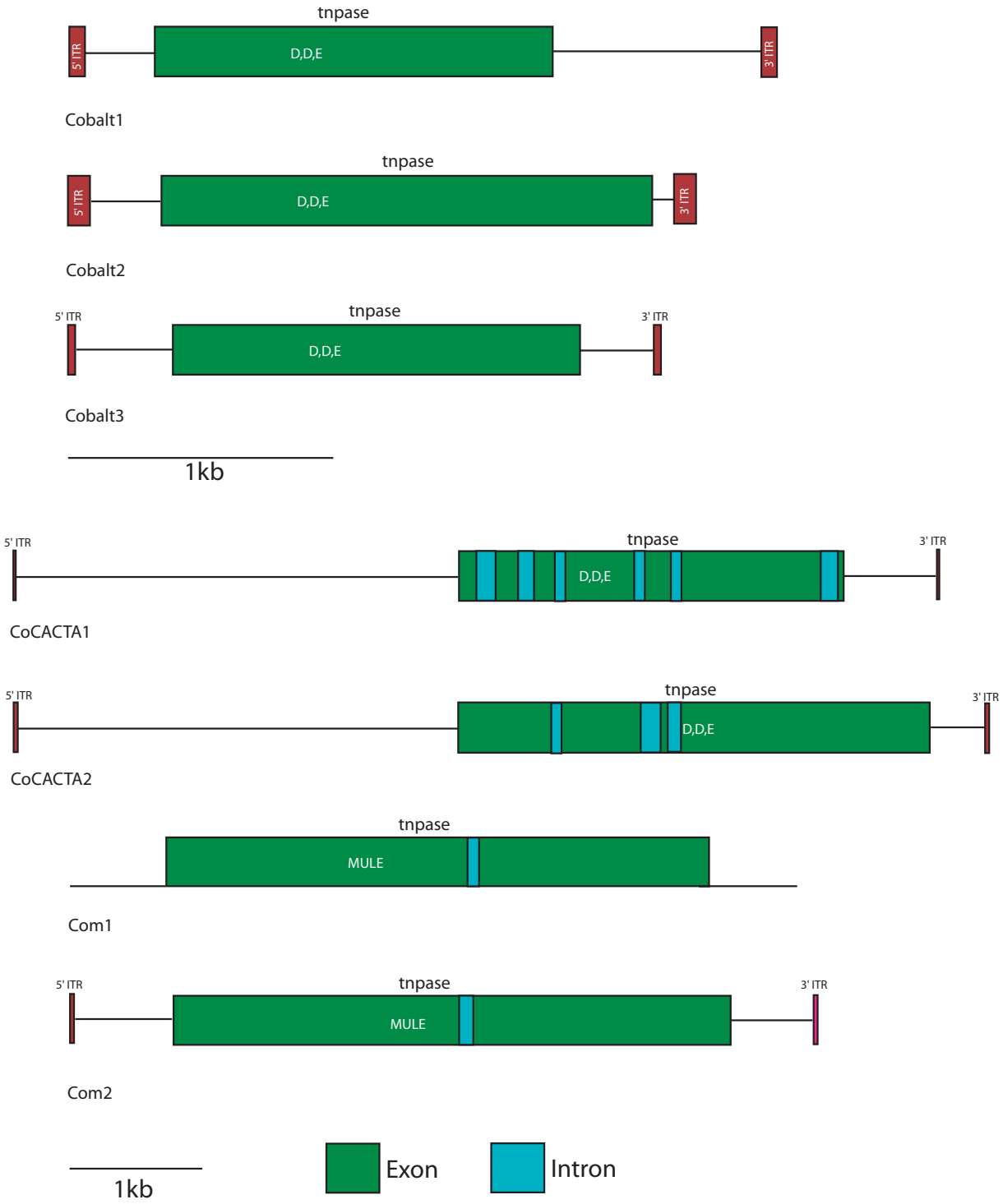
Supplementary Information

Supplementary Figures

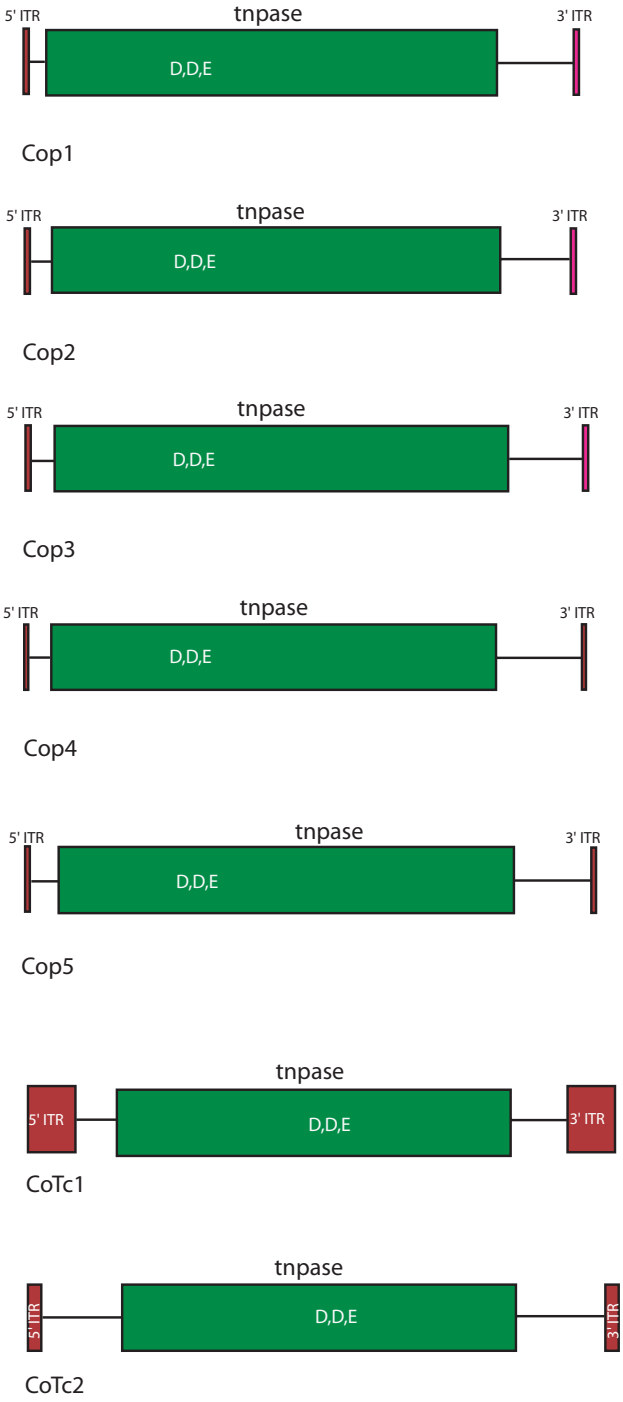


Supplementary Figure S1 continued

(C)

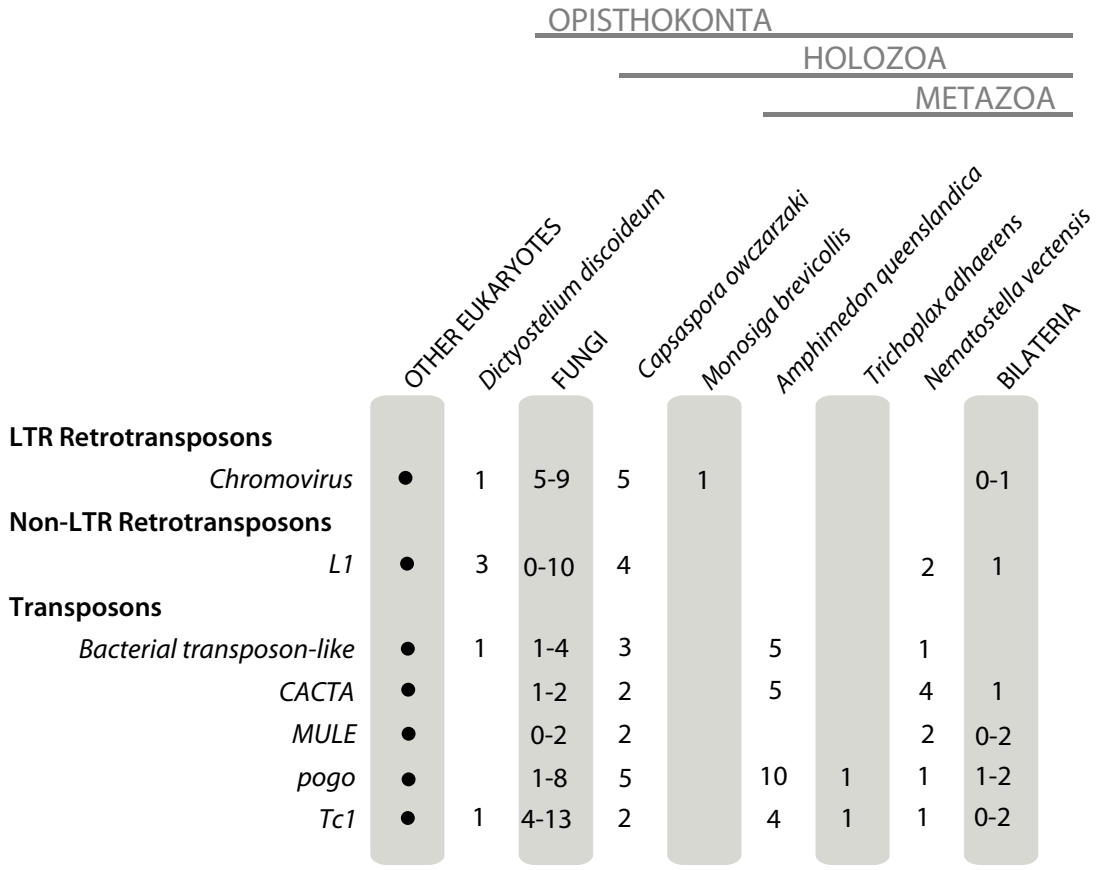


Supplementary Figure S1 continued



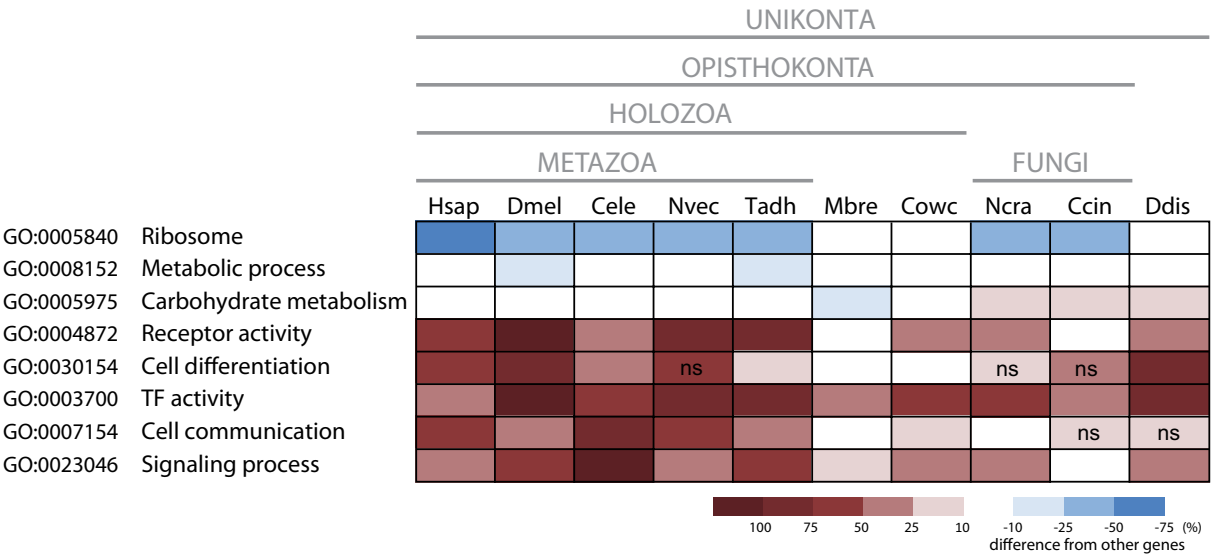
Supplementary Figure S1. Genomic organization of the 23 families of transposable element characterized in the *C. owczarzaki* genome

(A) LTR retrotransposons: Red boxes represent long terminal repeat sequences, green boxes represent *gag* open-reading frames (ORFs), dark blue boxes represent *pol* ORFs and light blue boxes represent *gag+pol* polyprotein ORFs. Protein coding domains are indicated as follows: CCHC, RNA binding motif; CD, chromodomain; IN, integrase; P, protease; RT, reverse transcriptase. Non-coding regions are indicated as follows: LTR, long terminal repeat. (B) Non-LTR retrotransposons: Green boxes represent *gag* open-reading frames (ORFs) and dark blue boxes represent *pol* ORFs. Protein coding domains are indicated as follows: CCHC, RNA binding motif; ENDO, endonuclease domain; RT, reverse transcriptase. Non-coding regions are indicated as follows: UTR, untranslated region. (C) Transposons: Red boxes represent inverted terminal repeat sequences, green boxes represent *tnpase* exon sequences, and light blue boxes represent *tnpase* intron sequences. Protein coding domains are indicated as follows: D,D,E, aspartic acid and glutamic acid catalytic domain; MULE, *Mutator-like element* transposase domain. Non-coding regions are indicated as follows: ITR, inverted terminal repeat.



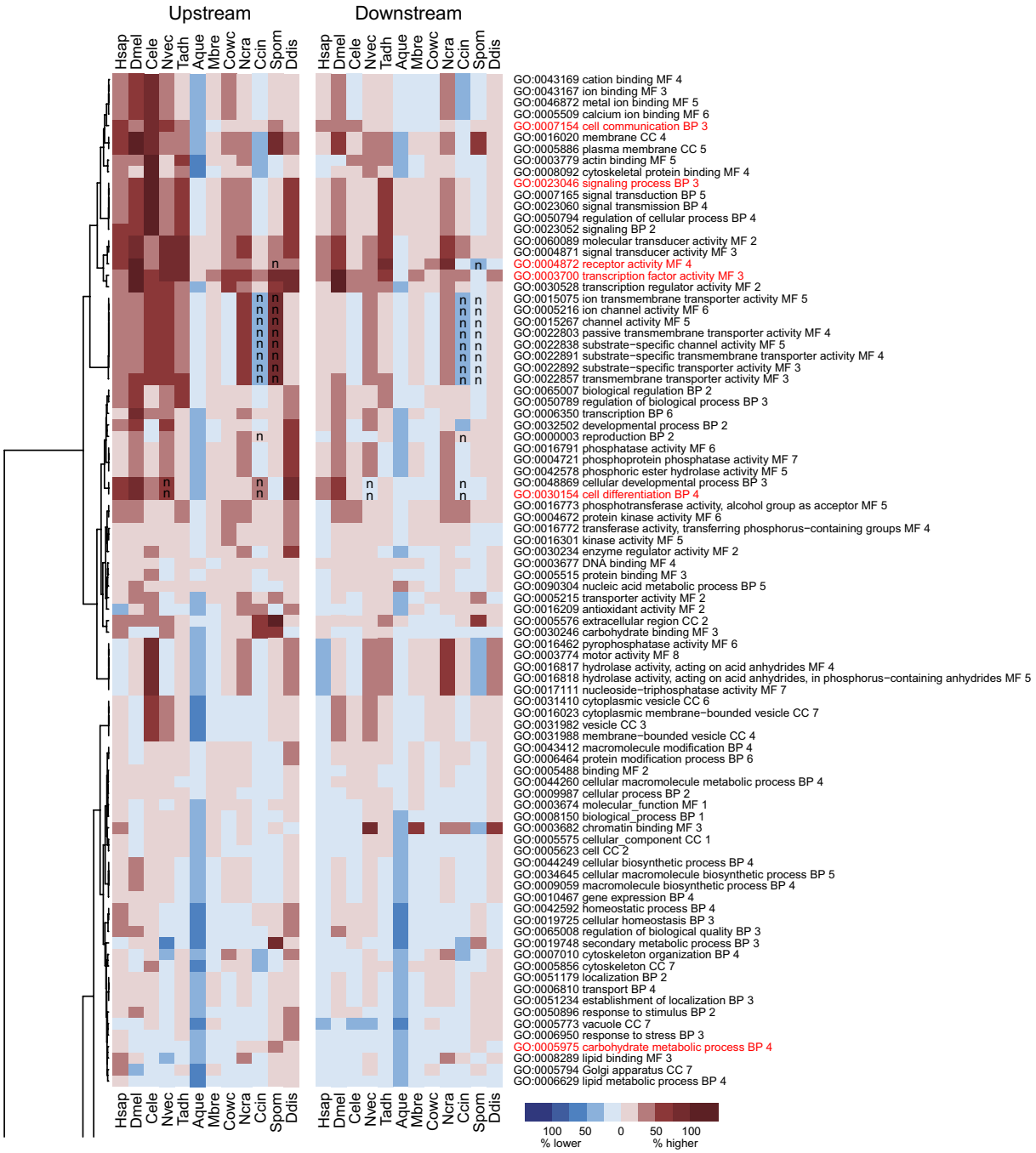
Supplementary Figure S2. Presence of transposable element superfamilies found in the *C. owczarzaki* genome in other eukaryotic genomes

One superfamily of LTR retrotransposons, one superfamily of non-LTR retrotransposons, and five superfamilies of canonical transposons were searched for in eukaryotic genomes. Family numbers (see text for the definition) are presented. The number of copies derived from each transposable element family in the *C. owczarzaki* genome is shown in Supplementary Table S1.

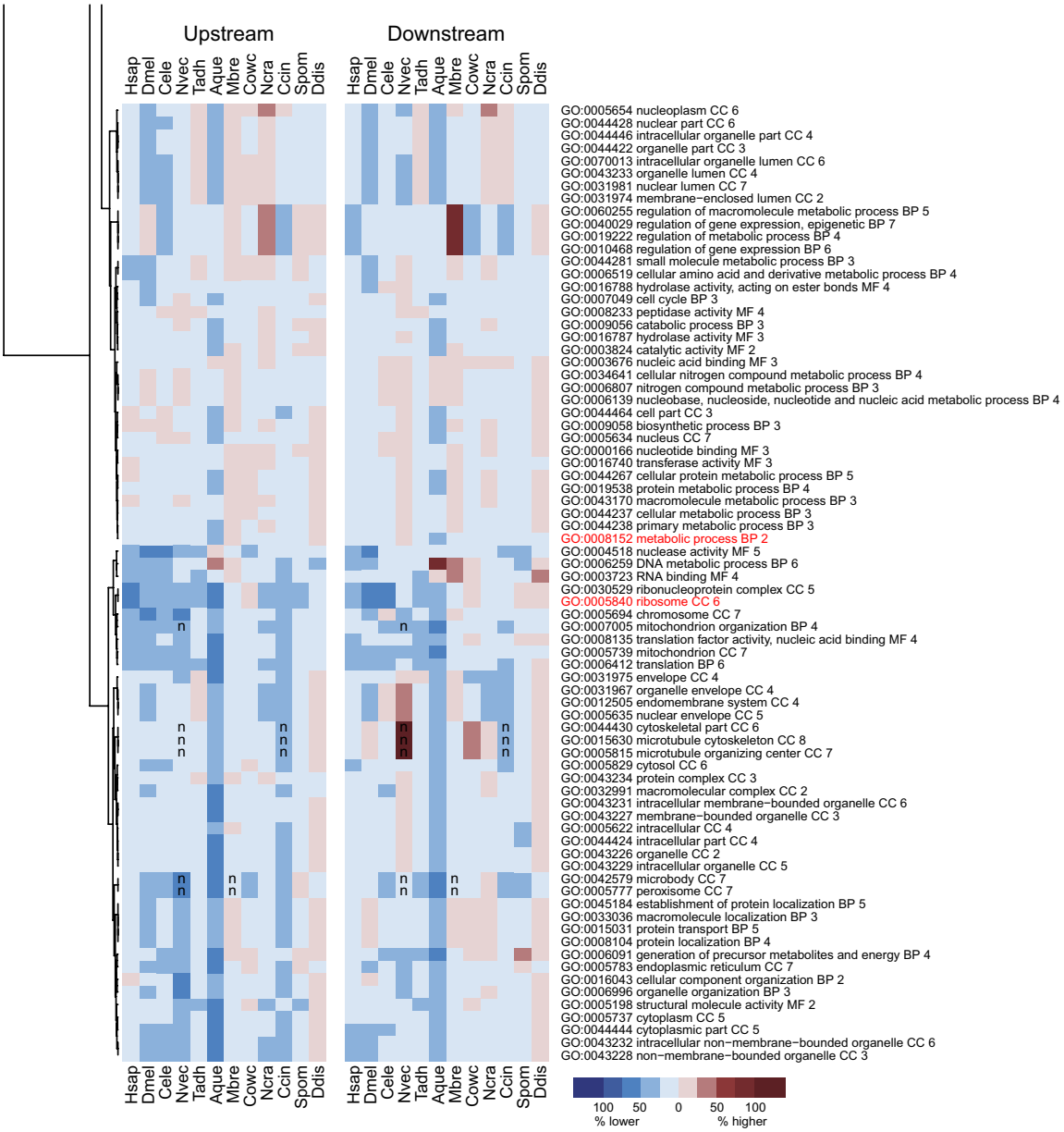


Supplementary Figure S3. Intergenic distance in unikonts

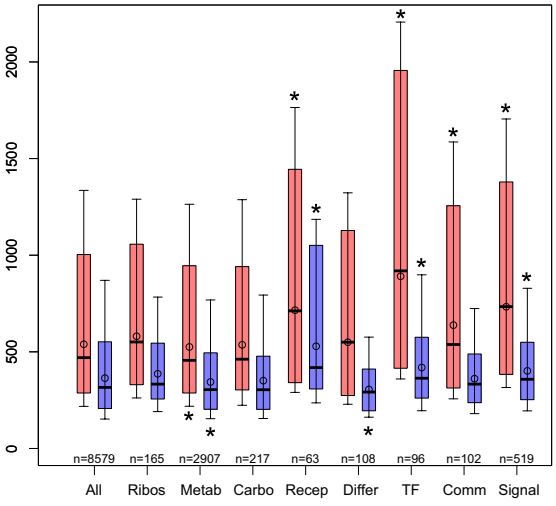
The geometric means of upstream intergenic lengths of protein-coding sequences in eight GO term categories relative to that of all other genes are displayed by color code. The cell is left blank when the difference is less than 10%. The detailed data are in Supplementary Figures S4 and S5). ns, not statistically significant though a more than 10% difference from other genes is observed (t-test p≥0.05). Hsap, *H. sapiens*; Dmel, *D. melanogaster*; Cele, *C. elegans*; Nve, *N. vectensis*; Tadh, *T. adhaerens*; Mbre, *M. brevicollis*; Cowc, *C. owczarzaki*; Ncra, *N. crassa*; Ccin, *C. cinerea*; Ddis, *D. discoideum*.



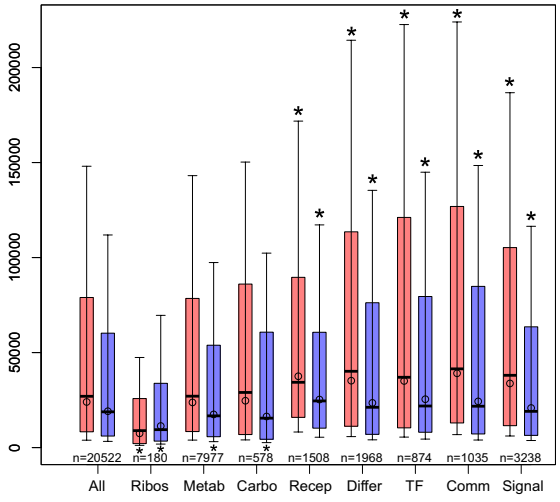
Supplementary Figure S4 continued



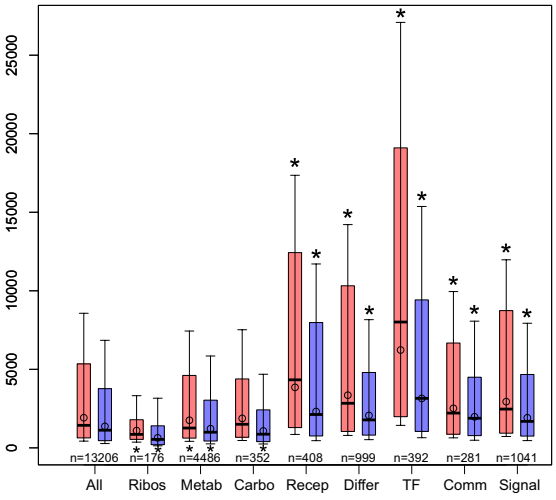
Supplementary Figure S4. Intergenic region sizes of genes in GO categories
A geometric mean of intergenic lengths of genes belonging to each GO category relative to that of all other genes is plotted in a heatmap. The top-level categories and the GO levels within them are shown at the end of GO names (BC, biological process; CC, cellular component; MF, molecular function). GOs lacking in any of the analyzed 12 species (Hsap, *H. sapiens*; Dmel, *D. melanogaster*; Cele, *C. elegans*; Nvec, *N. vectensis*; Tadh, *T. adhaerens*; Aque, *A. queenslandica*; Mbre, *M. brevicollis*; Cowc, *C. owczarzaki*; Ncra, *N. crassa*; Ccin, *C. cinereus*; Spom, *S. pombe*; Ddis, *D. discoideum*) are not shown. The upstream and downstream analyses are shown on the left and right, respectively. Cells with less than 10 genes are labelled with the letter n. GO names shown in Supplementary Figure S3 are in red. The dendrogram on the left was calculated for the upstream values.



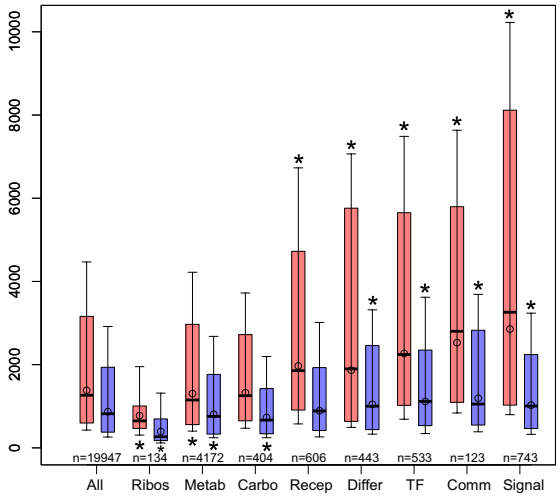
C. owczarzaki



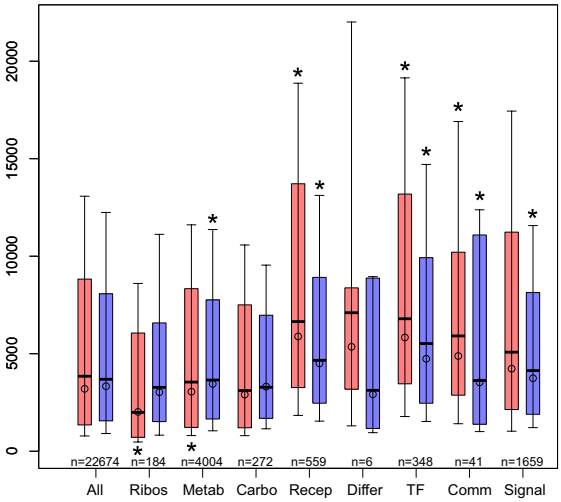
H. sapiens



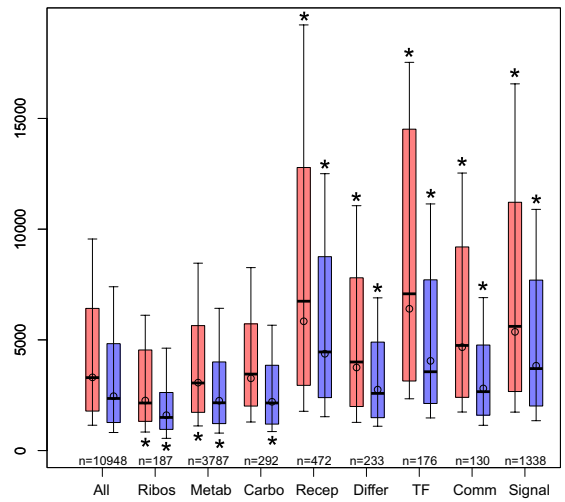
D. melanogaster



C. elegans

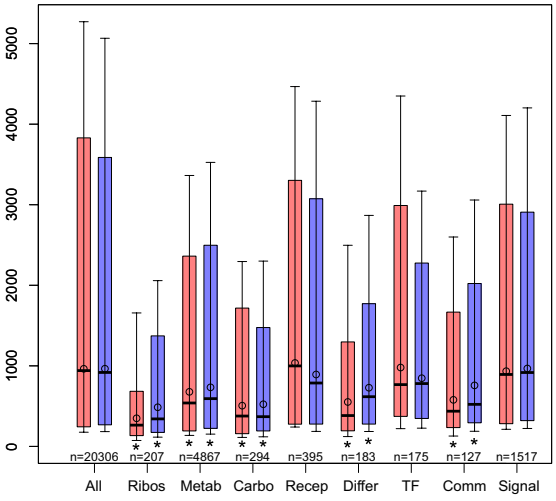


N. vectensis

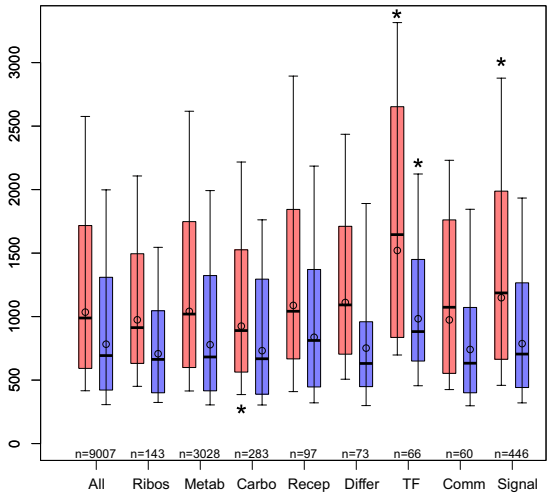


T. adhaerens

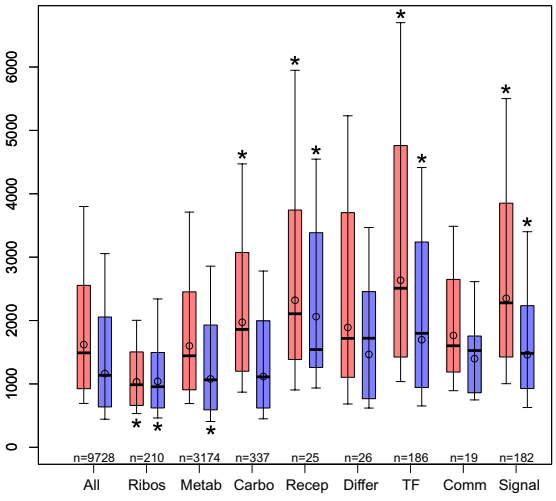
Supplementary Figure S5 continued



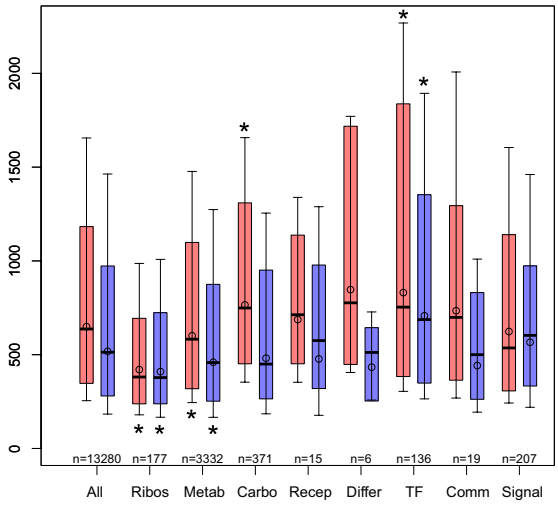
A. queenslandica



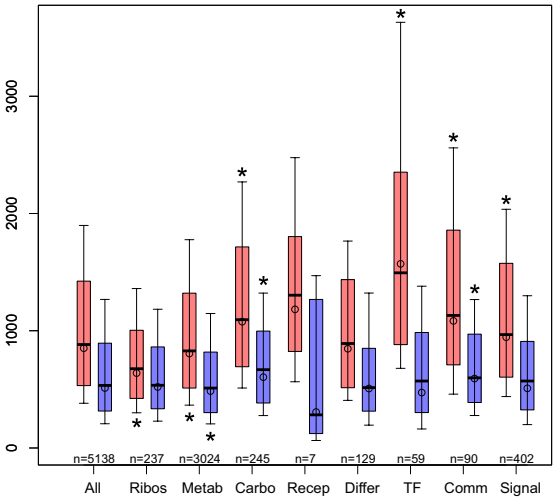
M. brevicollis



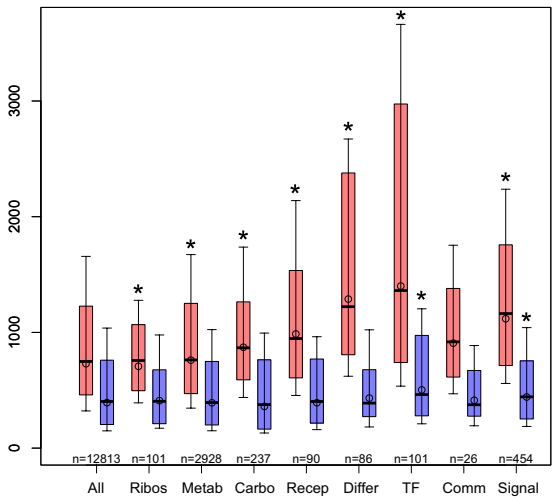
N. crassa



C. cinereus



S. pombe

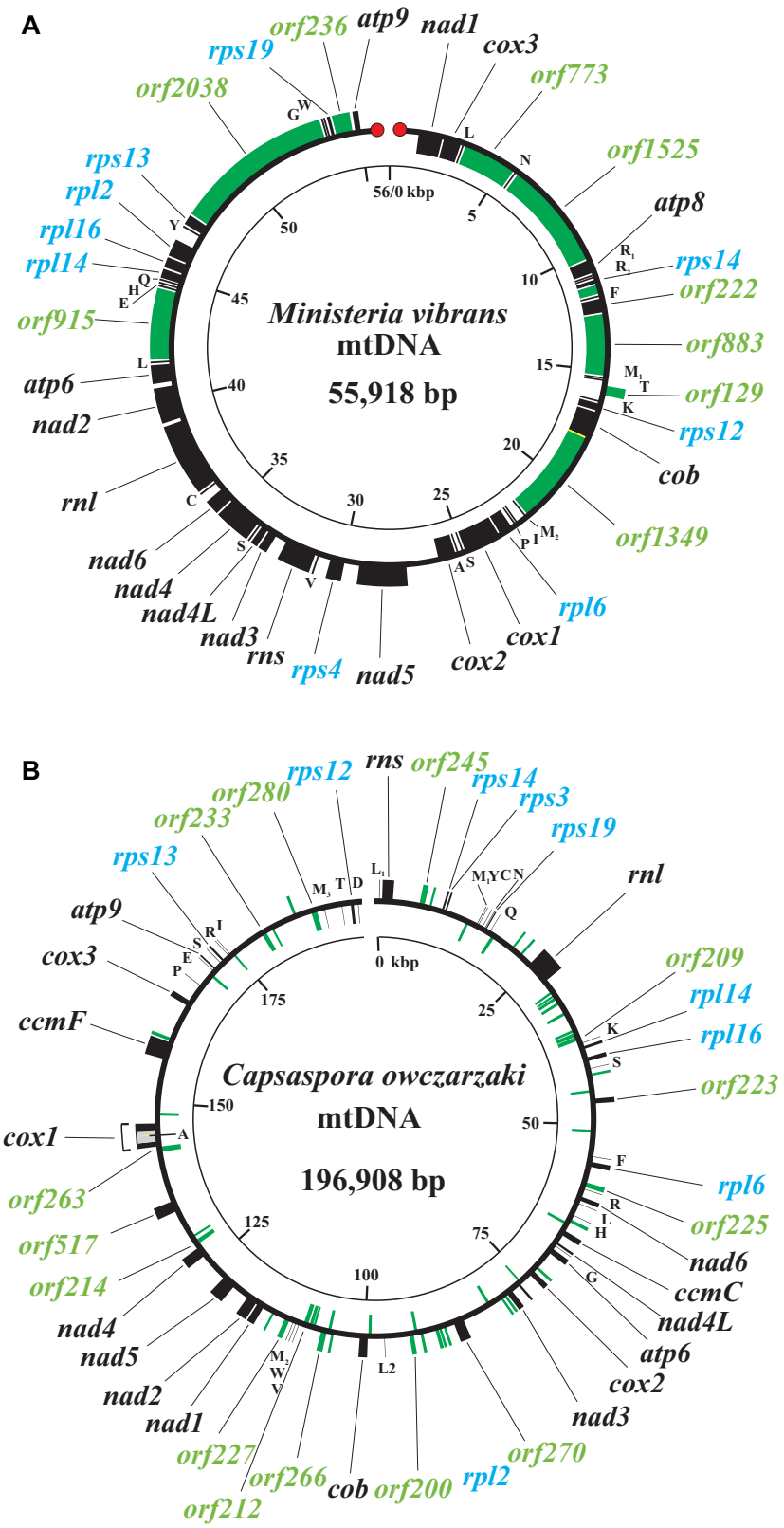


D. discoideum

Supplementary Figure S5 continued

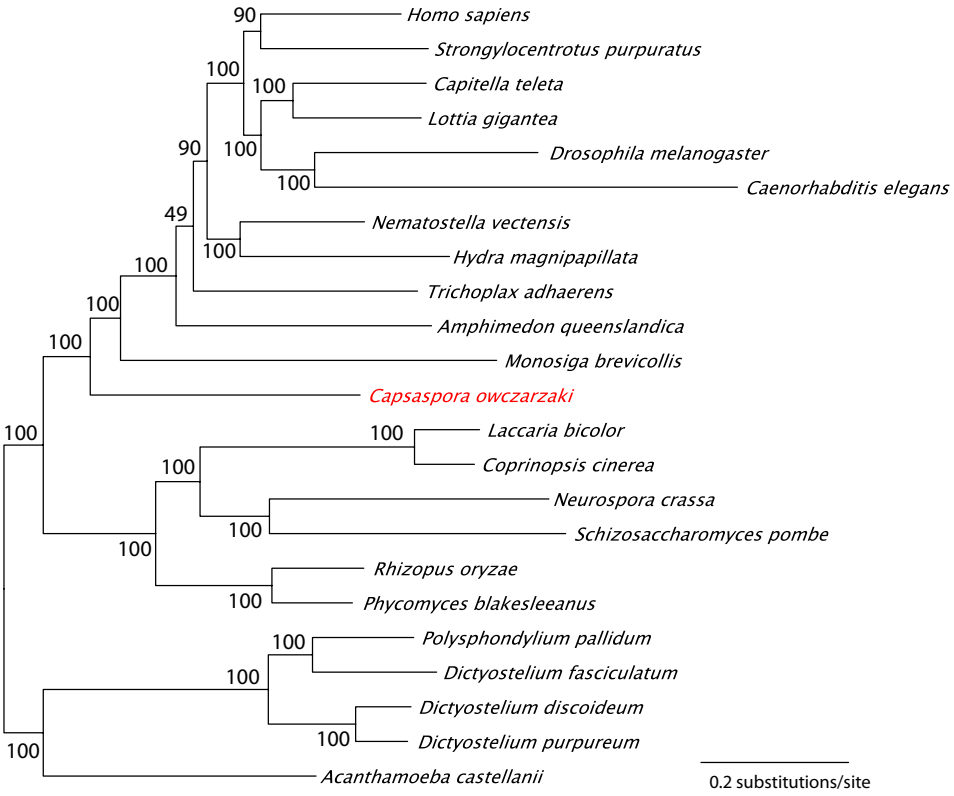
Supplementary Figure S5. Intergenic lengths of selected GO categories in 12 species

Lengths of intergenic regions are plotted for eight selected GO categories: Ribo, ribosome (GO:0005840); Metab, metabolic process (GO:0008152); Carbo, carbohydrate metabolic process (GO:0005975); Recep, receptor activity (GO:0004872); Differ, cell differentiation (GO:0030154); TF, transcription factor activity (GO:0003700); Comm, cell communication (GO:0007154); Signal, signaling process (GO:0023046). Box plot indicates the lower and upper quartiles with the median. Red and blue boxes represent the upstream and downstream analyses, respectively. Circles and error bars represent the geometric means ± standard deviation. The number of included genes for the upstream analysis is shown at the bottom. GO categories are labelled when the geometric mean is significantly (t-test $p<0.05$) larger (upper asterisk) or smaller (lower asterisk) than that of all other genes.

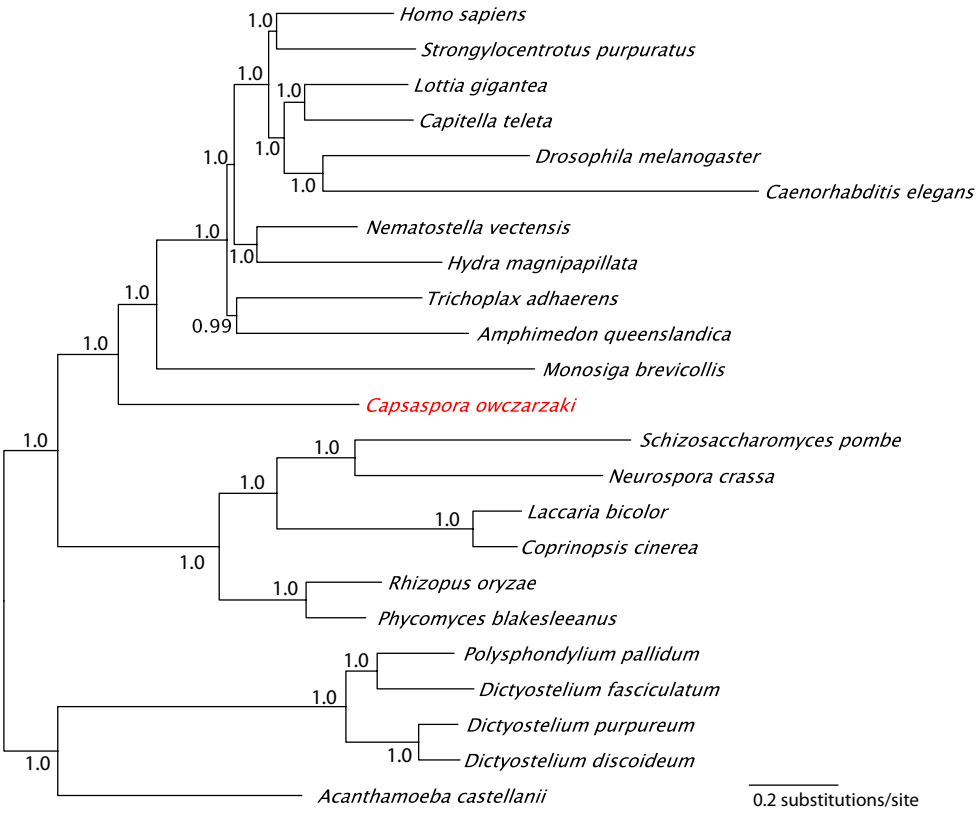


Supplementary Figure S6. Mitochondrial genetic maps of *M. vibrans* and *C. owczarzaki*.

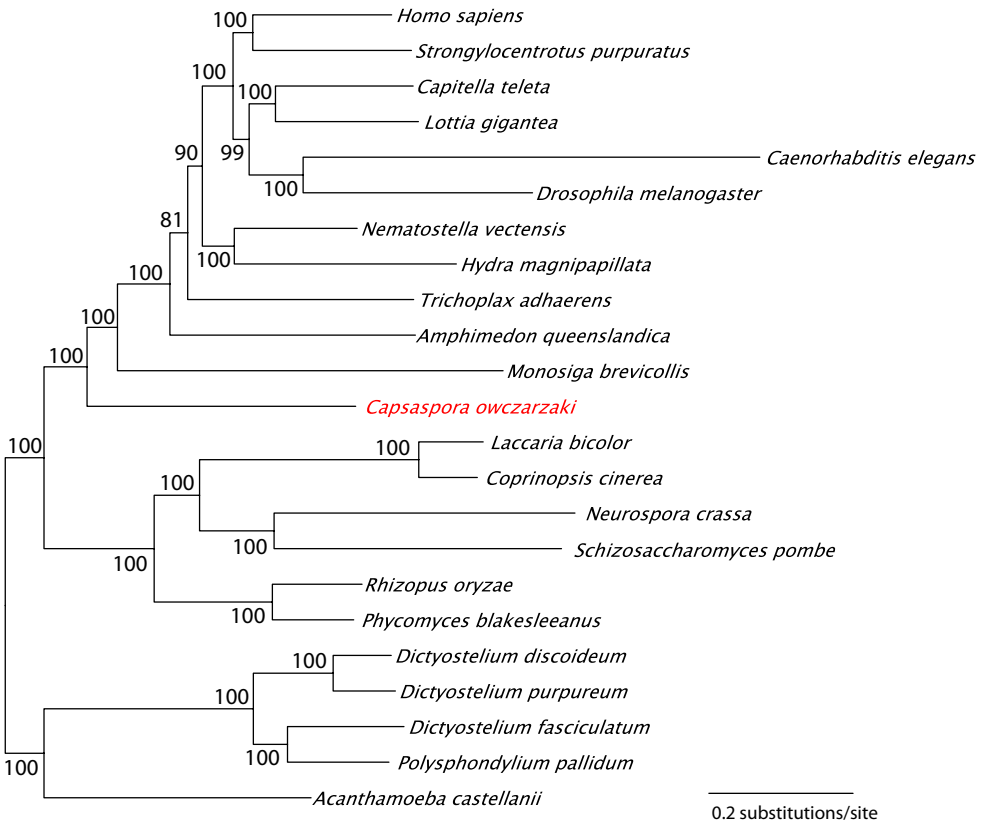
A and B, The linear M. *vibrans* mtDNA (A) has characteristic long inverted repeat sequences (extremities marked by red filled circles), whereas that of C. *owczarzaki* (B) is depicted linear but may have any other type of genome organization including the most common circular-mapping. Gene names for the standard set (black), ribosomal protein genes (blue) and ORFs (green, only larger than 200 a.a.) are indicated. Boxes of identified coding regions are filled black, ORFs green (including those between 100-200 a.a. in length), and introns light gray. The arc over *cox1* marks two exons interrupted by an intron. tRNA genes are labelled with capital letters, with the letter corresponding to the amino acid specified by the particular tRNA. Genes on the outer and inner circumference are transcribed in clockwise and counter-clockwise direction, respectively.



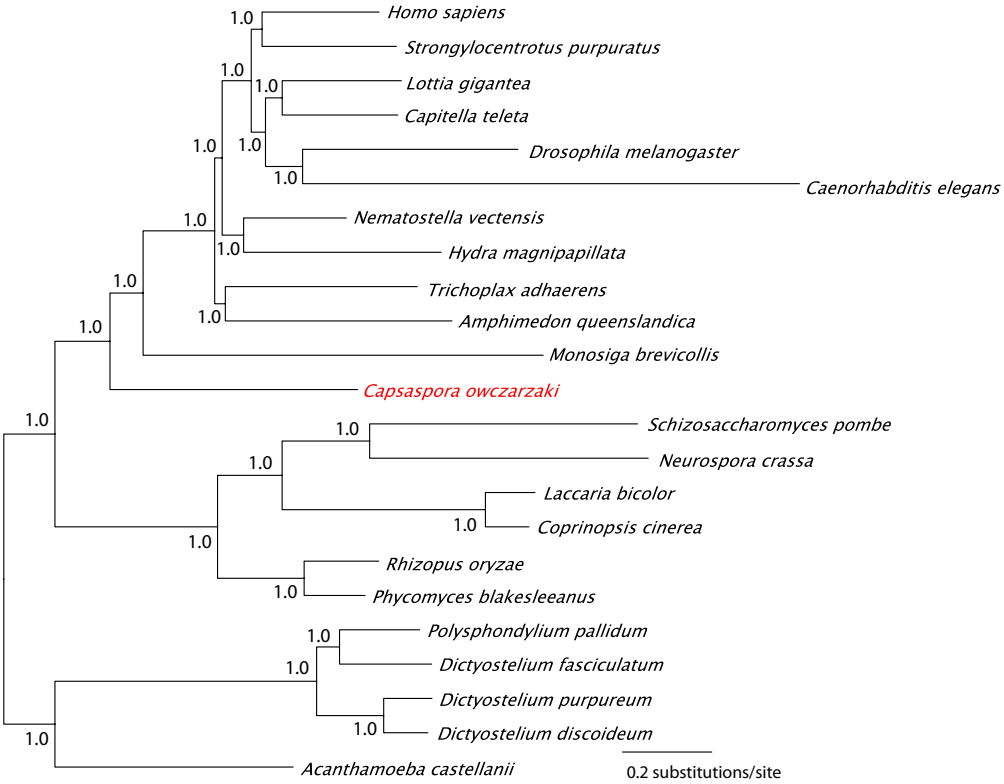
Supplementary Figure S7. fMBH-ML tree
The ML tree inferred from the fMBH dataset. *C. owczarzaki* in red. Bootstrap values are shown.



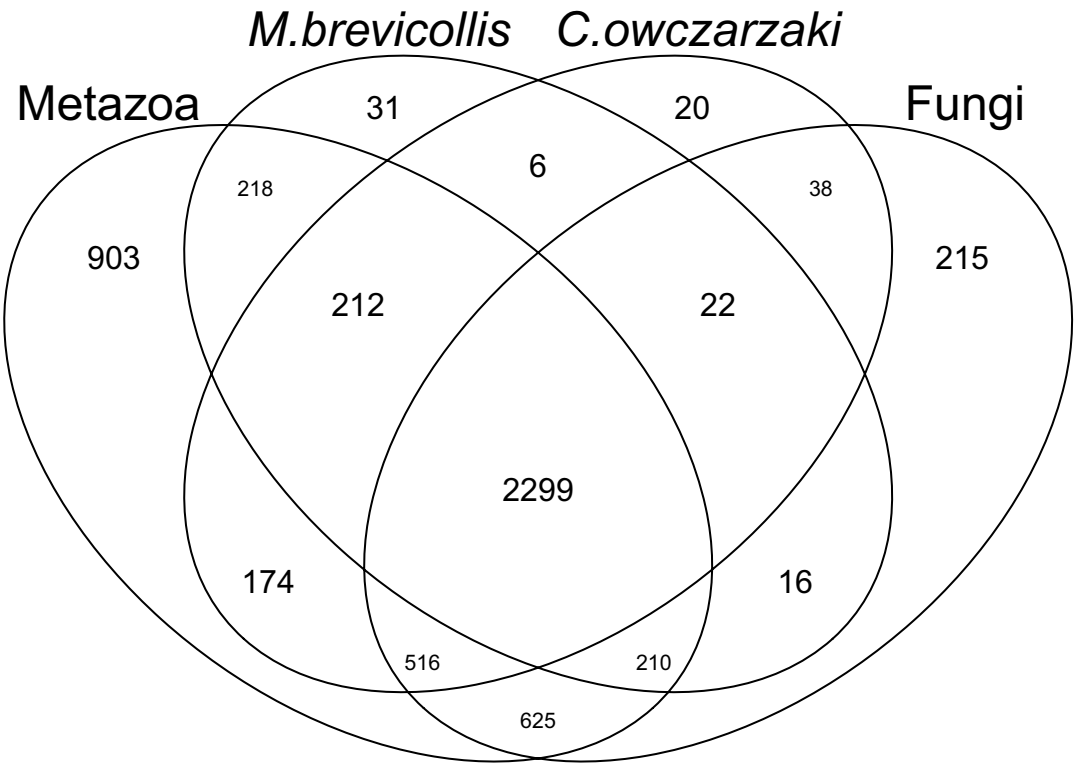
Supplementary Figure S8. fMBH-BI tree
The Bayesian tree inferred from the fMBH dataset. *C. owczarzaki* in red. Bayesian posterior probabilities are shown.



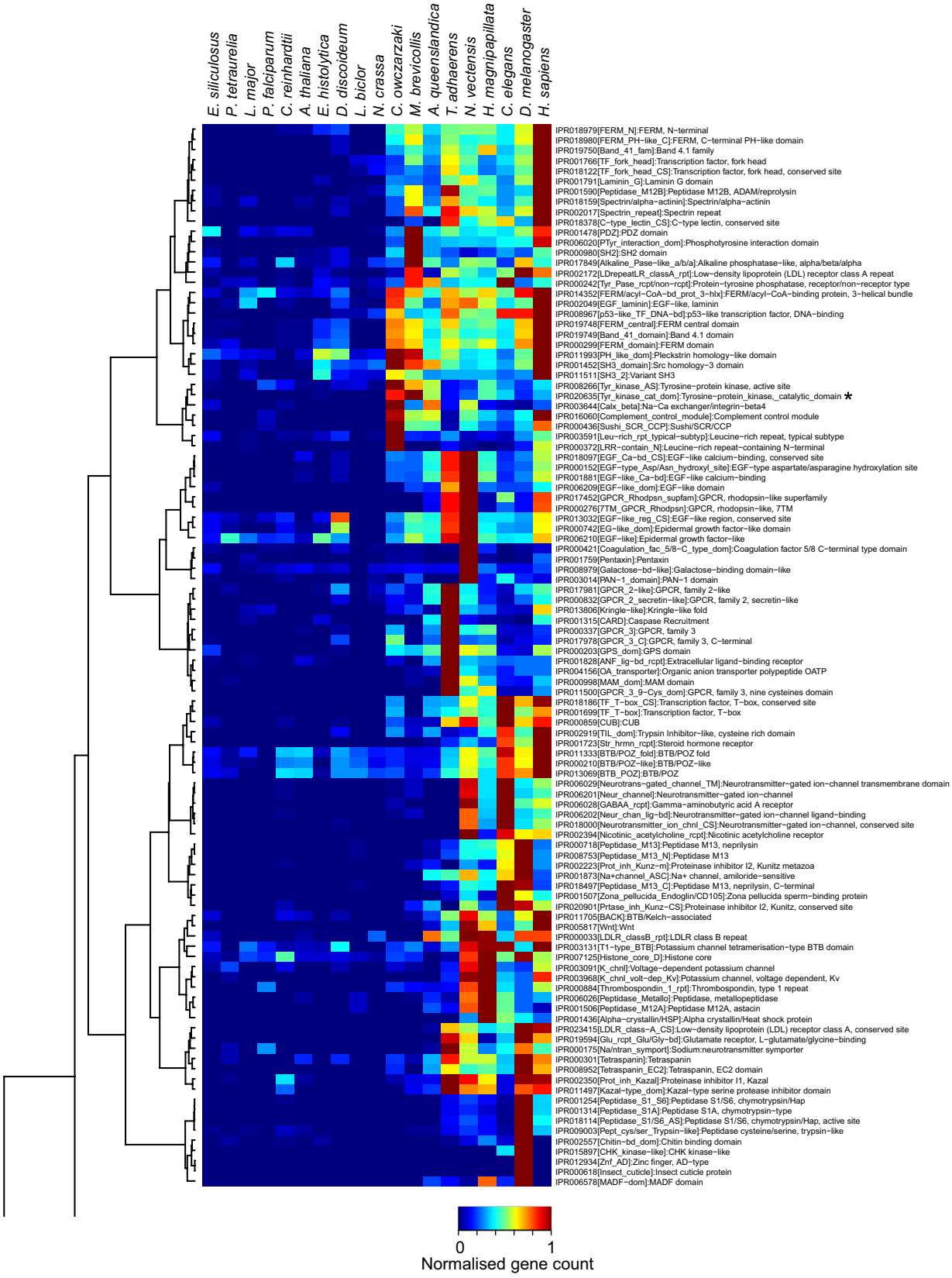
Supplementary Figure S9. 145POP-ML tree
The ML tree inferred from the 145POP dataset. *C. owczarzaki* in red. Bootstrap values are shown.



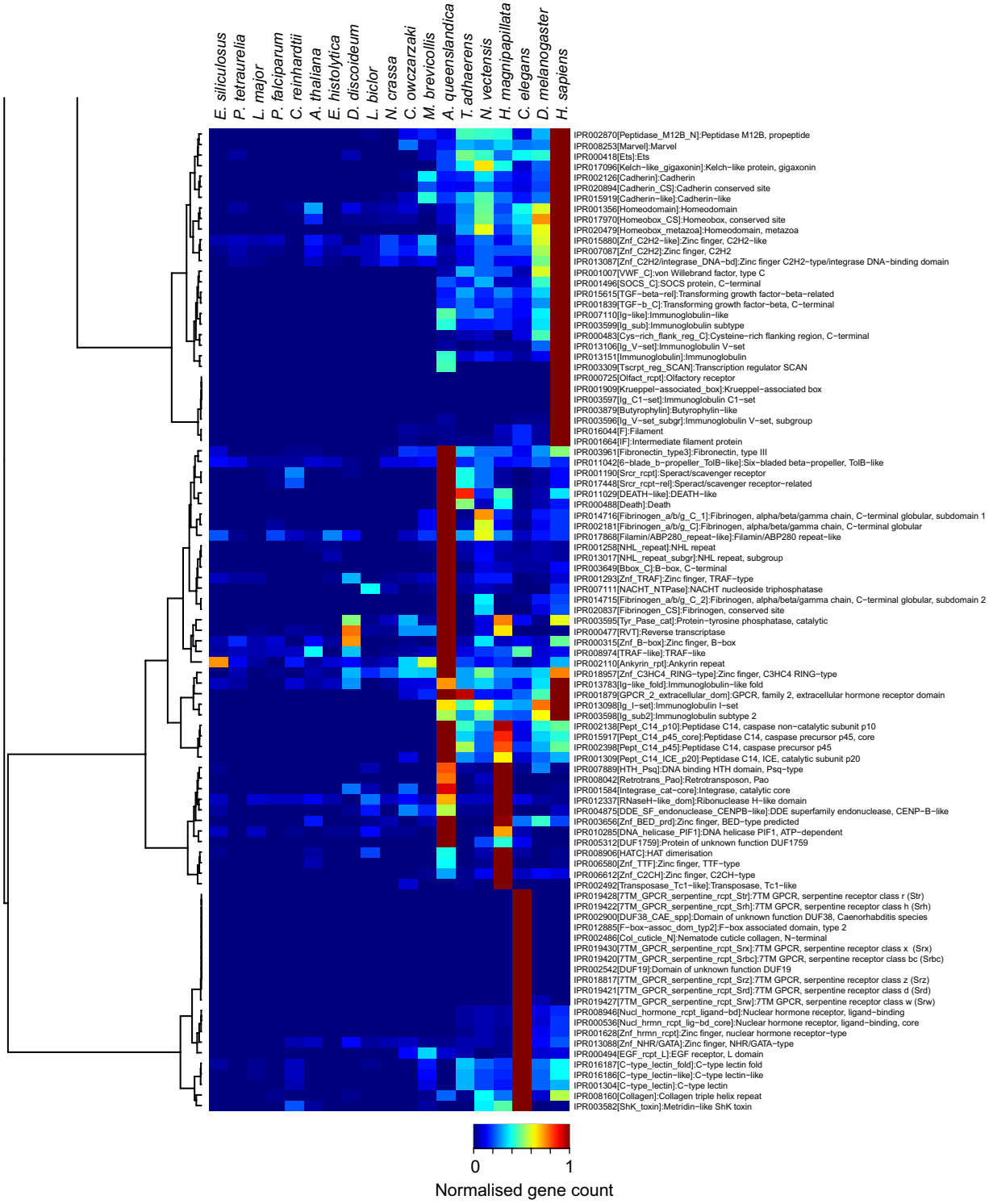
Supplementary Figure S10. fMBH-BI tree
The Bayesian tree inferred from the 145POP dataset. *C. owczarzaki* in red. Bayesian posterior probabilities are shown.



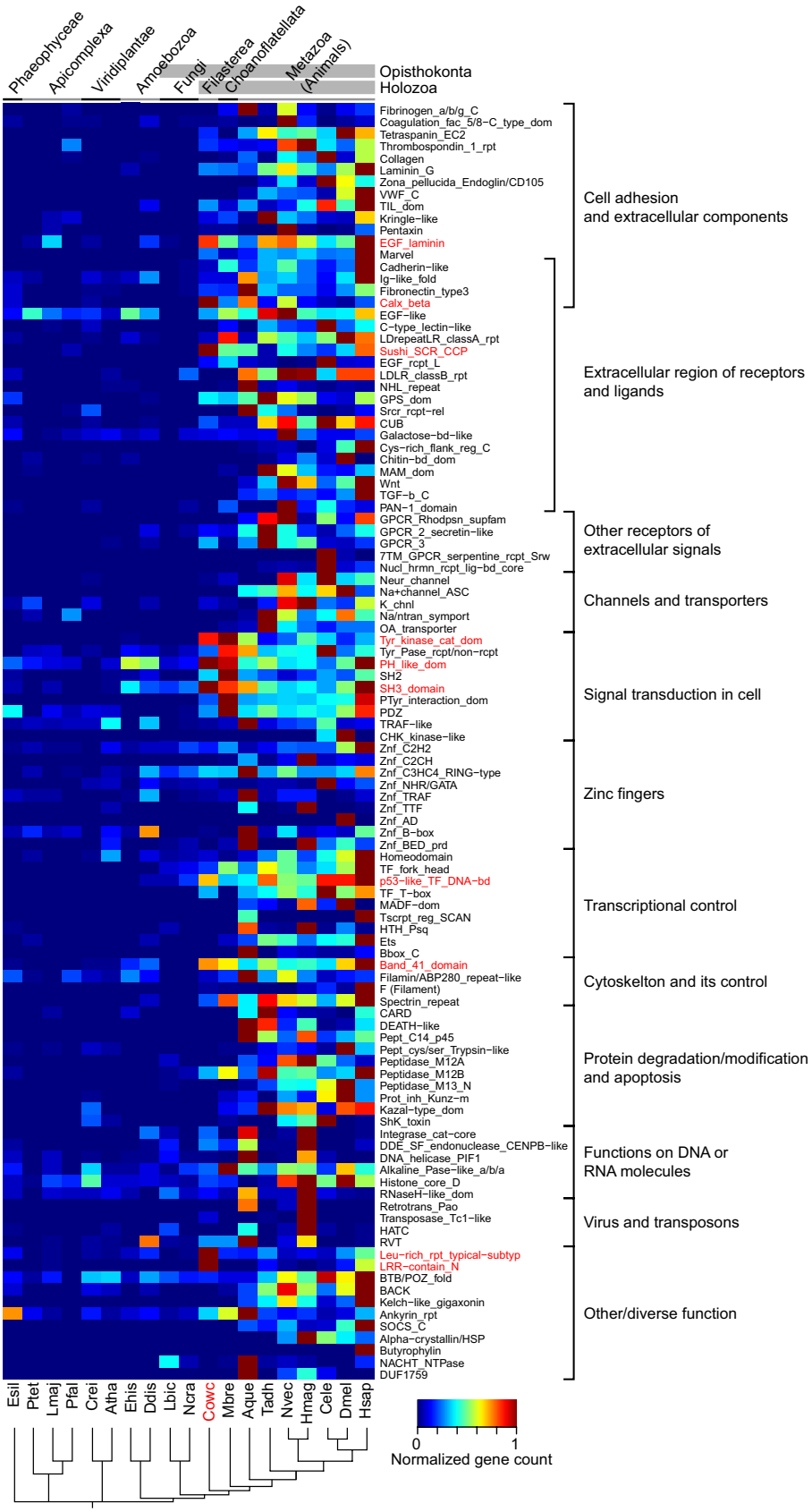
Supplementary Figure S11. Numbers of Pfam domains in the proteomes of four opisthokont lineages
The numbers of Pfam domans shared between, or unique to metazoans, *M. brevicollis*, *C. owczarzaki* and fungi are shown by a Venn diagram. See also Supplementary Table S7 for the detail.



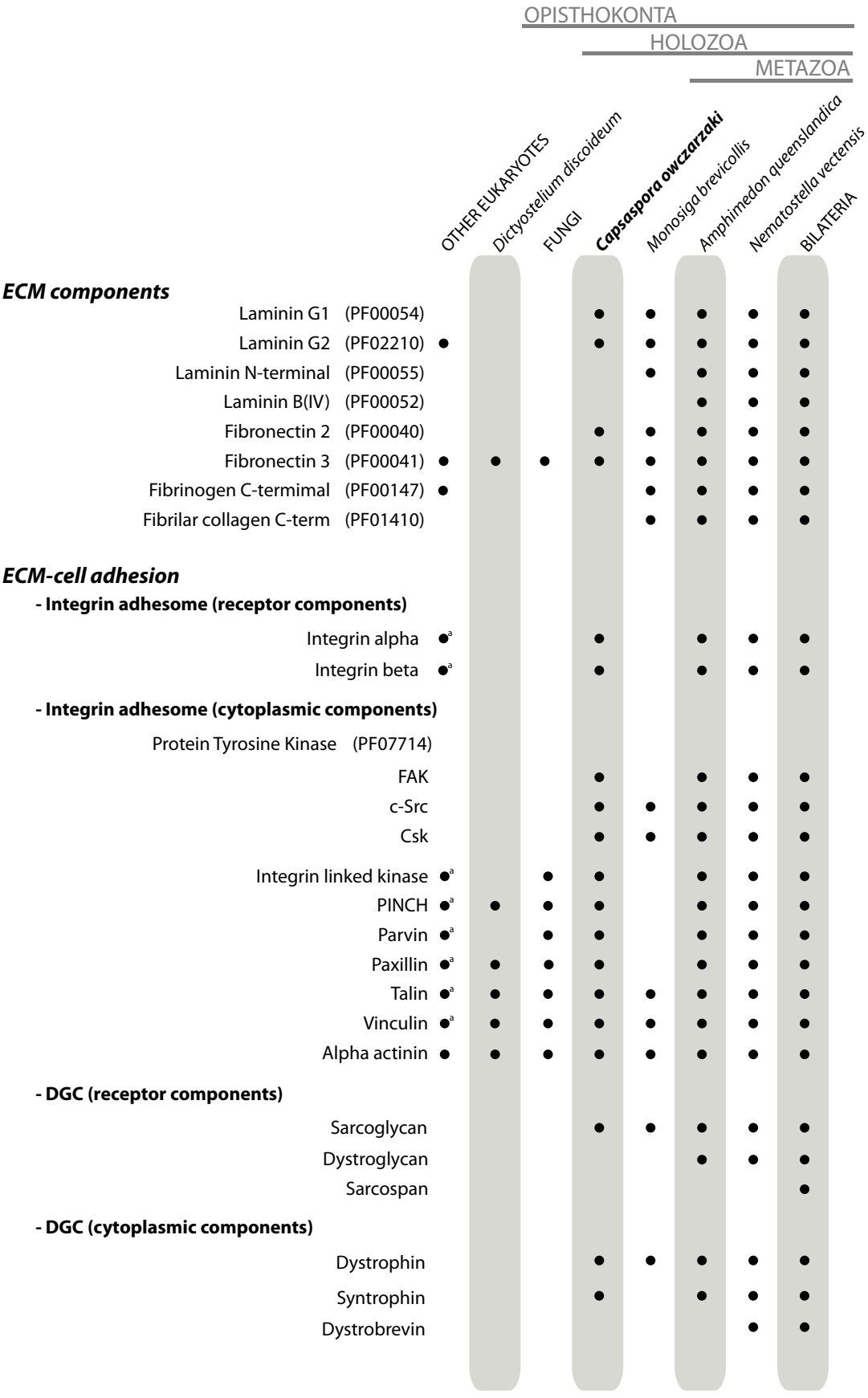
Supplementary Figure S12 continued



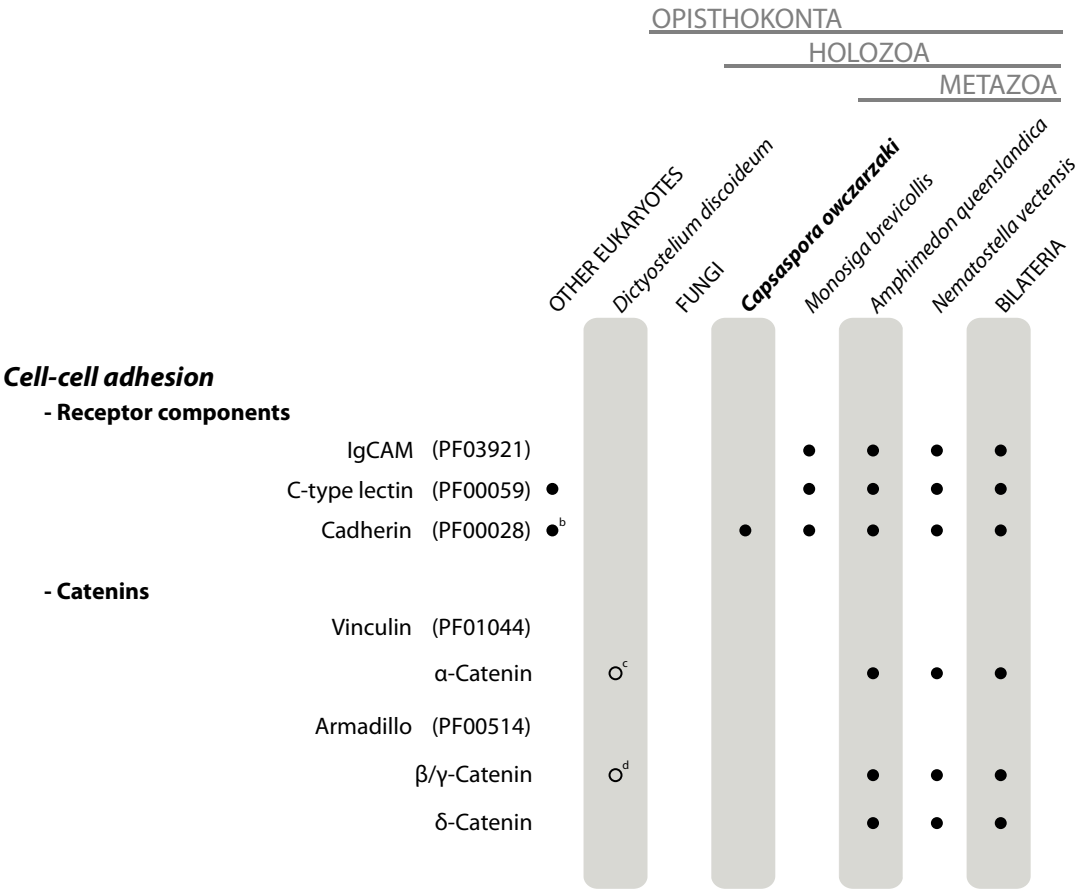
Supplementary Figure S12. Enrichment or depletion of protein domains in eukaryote genomes
All Interpro domains highly significantly enriched in metazoans ($p < 1.0e-20$) are shown, including the redundant ones. The gene counts for IPR020635 (asterisk) were entered manually. Interpro accession numbers, short names (in brackets) and full names are shown for each entry. A dendrogram on the basis of a clustering analysis is shown on the left**.**



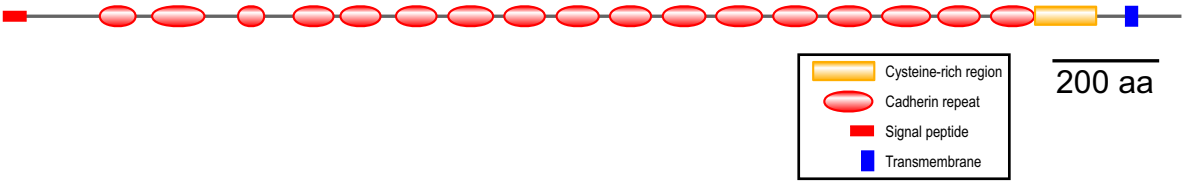
Supplementary Figure S13. Enrichment or depletion of protein domains in eukaryote genomes
We compressed the Interpro domains shown in Supplementary Figure S12 by removing redundant ones and single-taxon-specific ones. Selected domains were classified by their functional categories shown on the right. Domains with high relative gene counts (>0.65) in *C. owczarzaki* are in red. This is in principle the same figure with Figure 3 in the main text, but some functional categories abbreviated in Figure 3 are shown here.



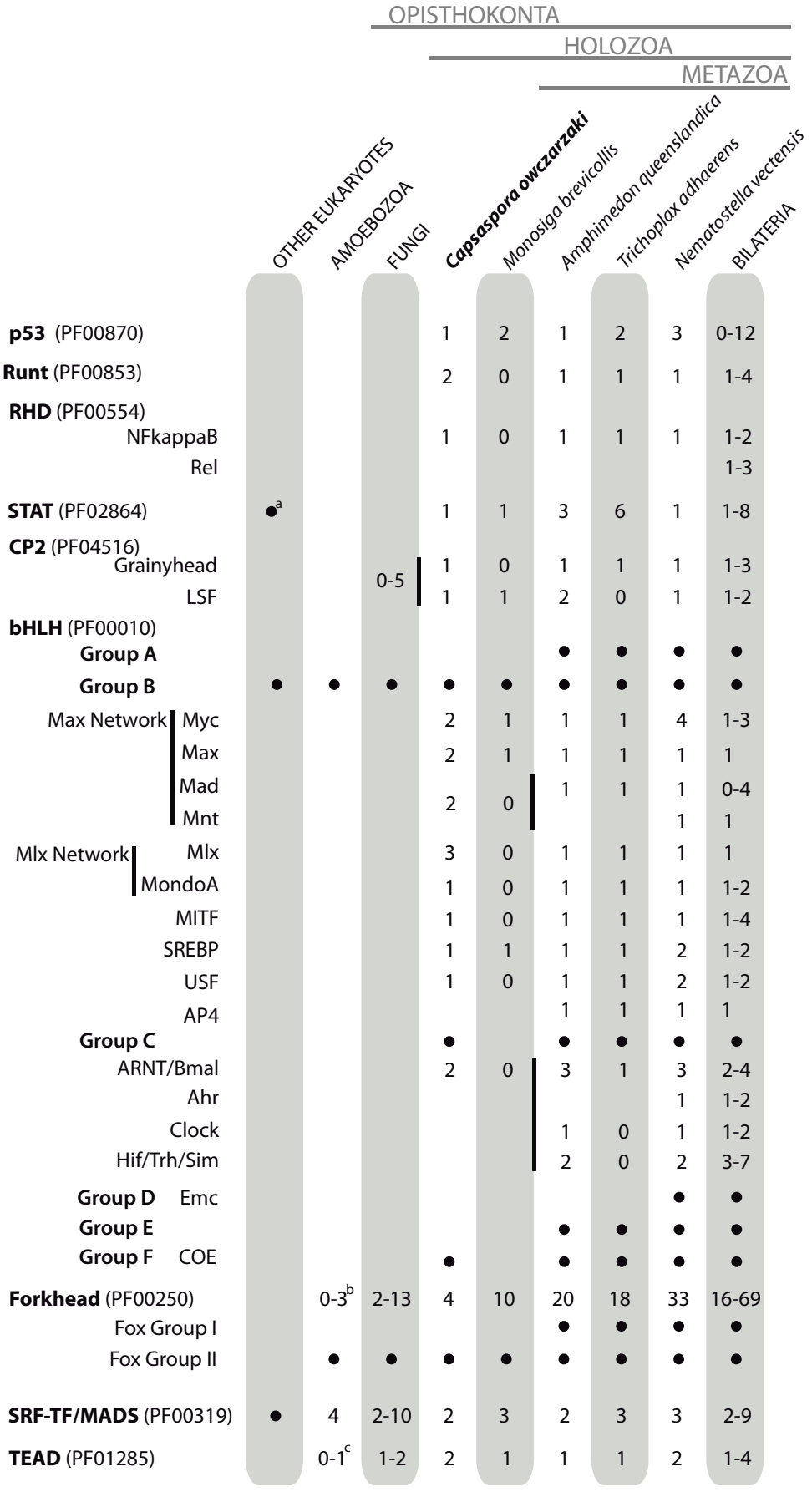
Supplementary Figure S14 continued



Supplementary Figure S14. Presence or absence of cell adhesion systems in *C. owczarzaki* and other eukaryotes
Pfam protein domains used for the HMMER search are shown with the Pfam accession numbers when a simple domain search was performed. Otherwise gene names are indicated. a, present in the apusozoan *T.trahens*[16]; b, present in the oomycetes *Pythium ultimum* and *Phytophtora infestans*[53]; c and d, experimentally suggested homologs[54,55].



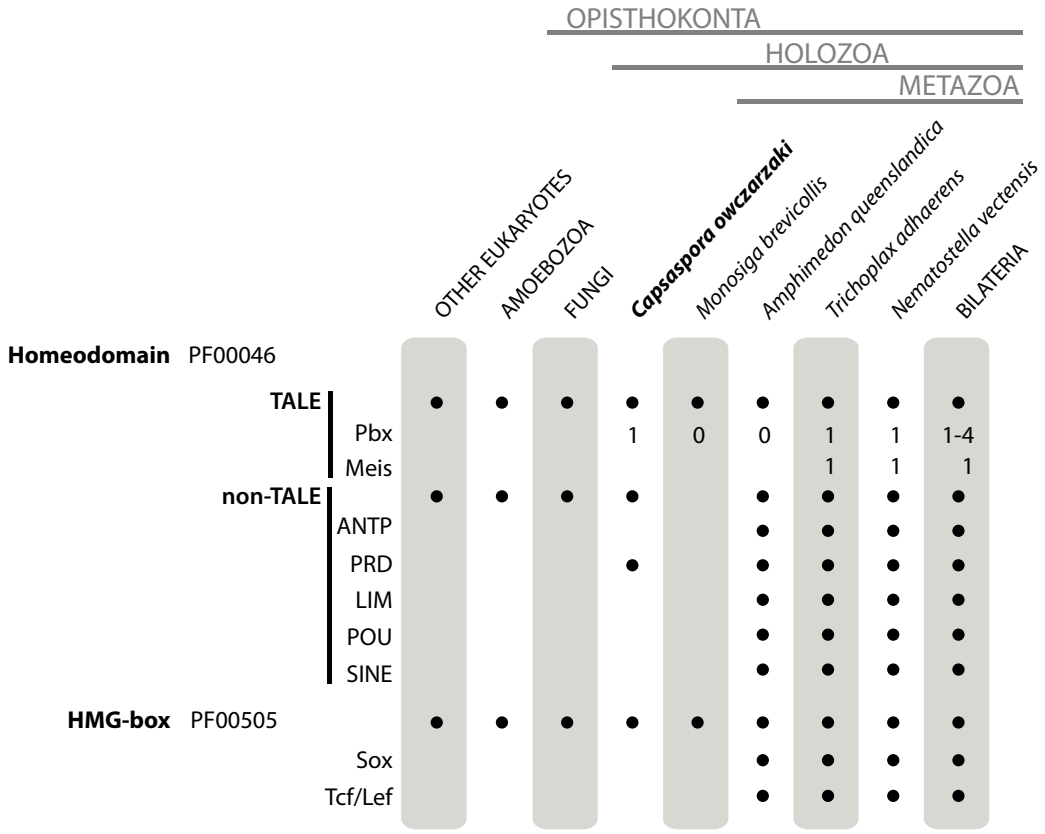
Supplementary Figure S15. *C. owczarzaki* receptor protein containing cadherin repeats
The only *C. owczarzaki* protein that contains cadherin repeats is schematically drawn. The cysteine-rich region and the cytoplasmic region cannot be reliably mapped to any known protein domain.



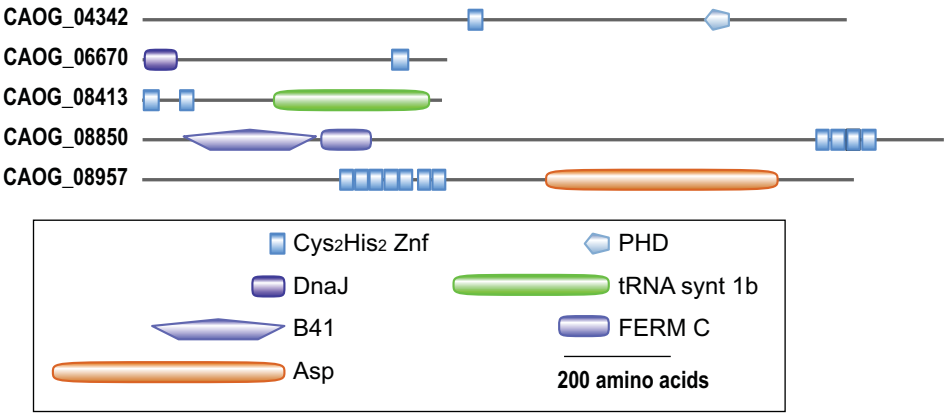
Supplementary Figure S16 continued

	OTHER EUKARYOTES	AMOEBOZOA	FUNGI	*Capsaspora owczarzaki*	*Monosiga brevicollis*	*Amphimedon queenslandica*	*Trichoplax adhaerens*	*Nematostella vectensis*	BILATERIA
	OPISTHOKONTA								
			HOLOZOA						
						METAZOA			
T-box (PF00907)			●[d]	●		●	●	●	●
Unclassified			0-1[d]	2	0	5			
Brachyury				1	0	0	1	1	1-2
Tbx1/15/20						1	0	6	3-6
Tbx4/5						1	0	1	1-2
Tbx2/3							1	1	1-2
Tbx6									2-3
Tbrain									1-3
bZIP (PF00170/07716)	●	●	●	●	●	●	●	●	●
Jun						1	1	3	1-2
Fos						1	0	4	1-3
XBP1						1	1	1	1
Atf3									1
MAF						1	1	1	1
BACH									1-2
Nfe2						2	1	3	1-2
B-ATF									1-2
PAR				1	0	1	0	4	1-3
C/EBP				2	0	2	2	2	1-2
Atf4/5					2	1	1	0	1-2
Oasis				1	2	2	0	2	1-2
Atf6				1	1	1	1	1	1
CREB				1	0	1	1	1	1-2
Atf2				1	1	1	1	1	1-3
Zinc Fingers									
Cys_2His_2 (PF00096)	●	19	33-128	42	74	54	52	226	117-718[e]
GATA (PF00320)	●	20	6-36	12	3	3	4	5	6-8
$Zn(II)_2Cys_6$ (PF00172)	●	2	21-90	9	0	0	0	0	0
CSL (PF09271)			0-7	1	1	1	1	1	1-2
RFX (PF02257)			0-1	1	3	3	4	5	2-7
HSF (PF00447)	●	1	4-14	3	2	1	2	3	1-6

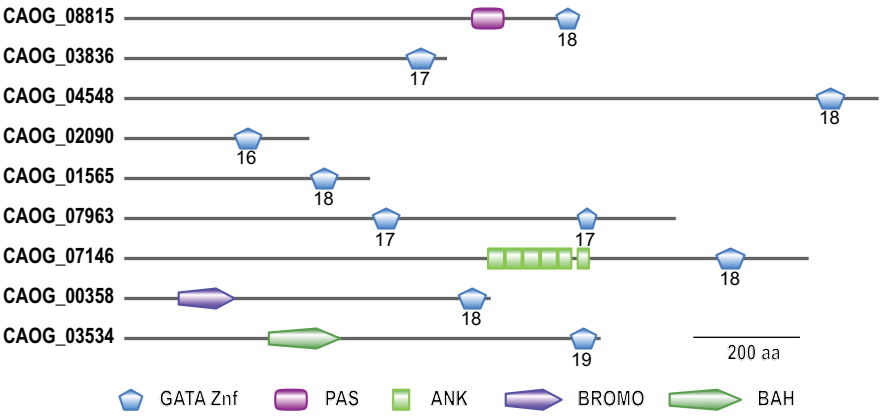
Supplementary Figure S16 continued



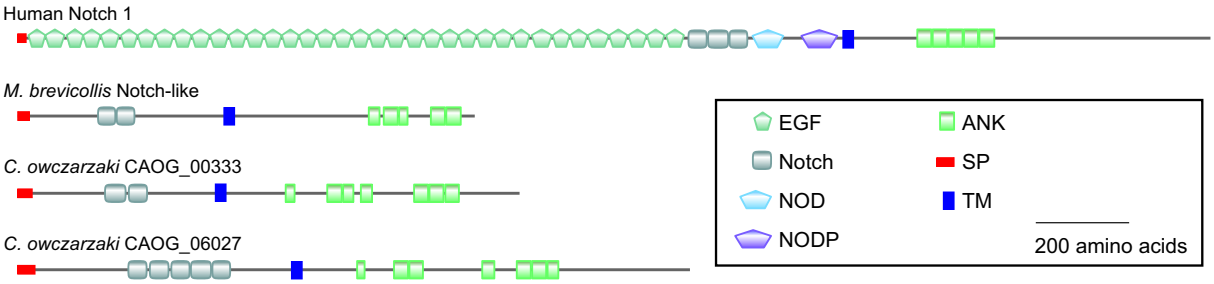
Supplementary Figure S16. Conservation of transcription factors
Presence of transcription factors (TFs) is indicated by a dot (or gene number, when their classification is reasonably robust). Pfam statistical models used for HMMER searches are shown after domain names. a, STAT domain present in *Thecamonas trahens*[17]; b, Forkhead domain present in *Acanthamoeba castellanii*[17]; c, TEA present in *Acanthamoeba castellanii*[19]; d, T-box present in *Spizellomyces punctatus*[17]; e, data taken from a previous publication[56].



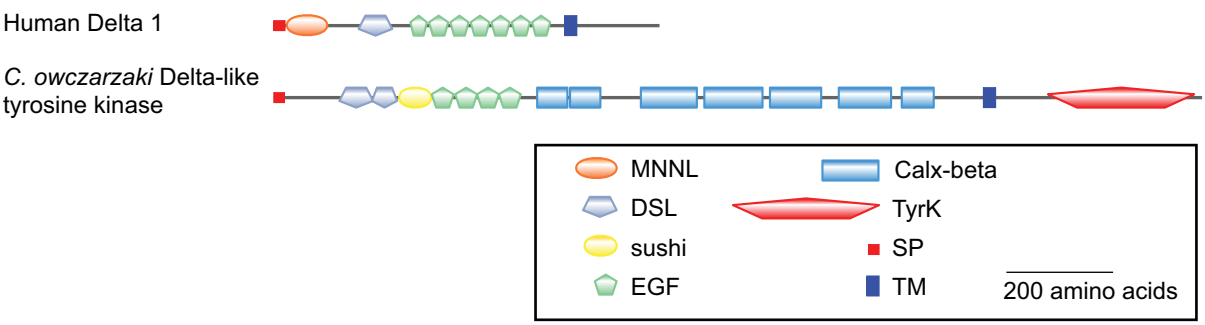
Supplementary Figure S17. Domain architectures of five *C. owczarzaki* Cys_2His_2 zinc fingers
Structures of the five *C. owczarzaki* Cys_2His_2 zinc finger proteins that have other protein domains than zinc fingers are schematically represented. Asp, aspartyl protease; B41, Band 4.1 homology; Cys_2His_2 Znf, Cys_2His_2 zinc finger; DnaJ, DnaJ molecular chaperone homology; FERM C, FERM C-terminal PH-like; PHD, plant homeodomain finger; tRNA synt 1b, tRNA synthetase class I (Trp and Tyr). Size of scheme is proportional to the actual sequence length.



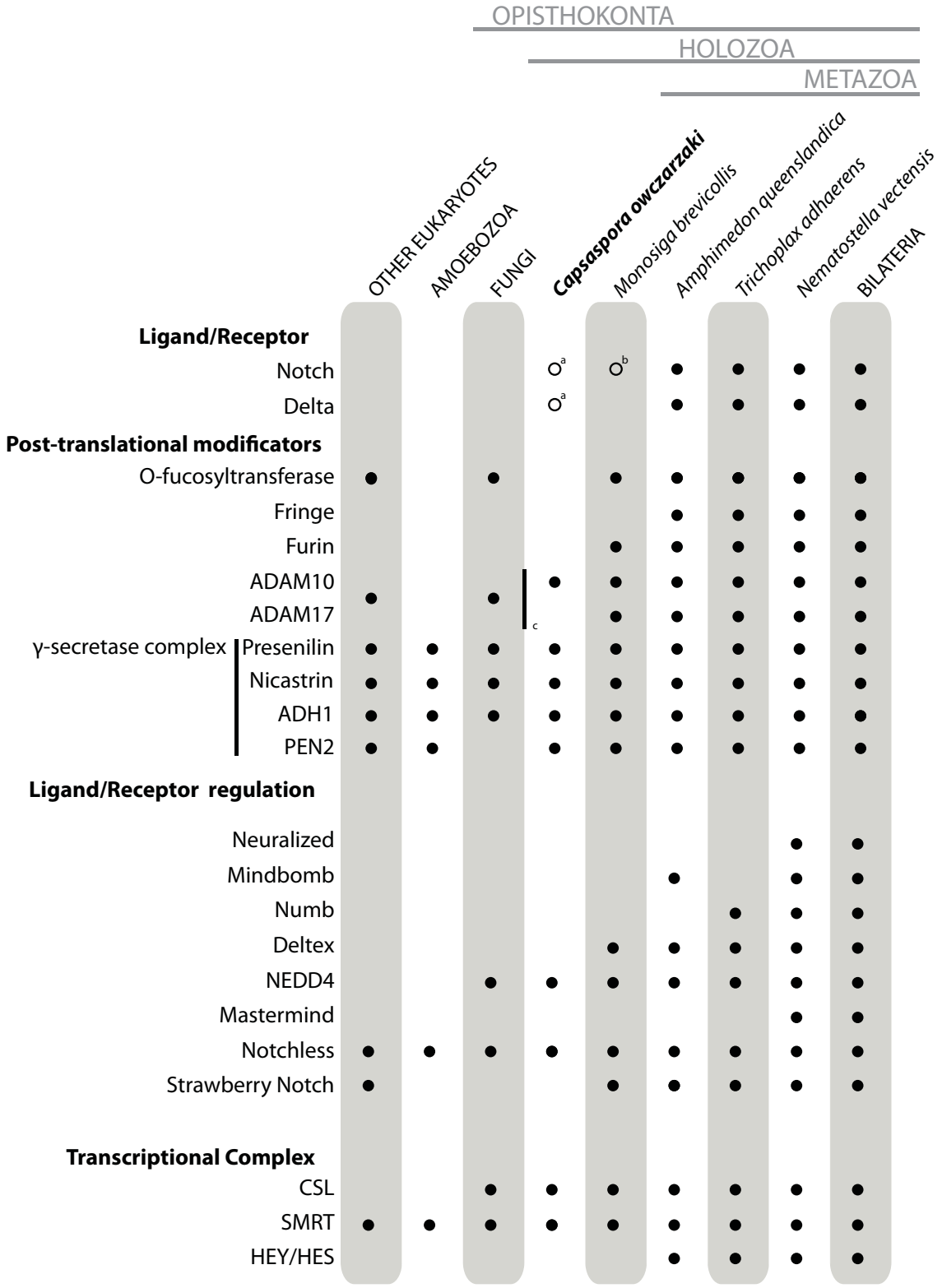
Supplementary Figure S18. Domain architectures of *C. owczarzaki* GATA factors
Structures of the nine *C. owczarzaki* GATA factors are schematically represented. Numbers of amino acids contained in the loop region is shown under the GATA zinc finger (GATA Znf) domains. ANK, Ankyrin repeat; BAH, bromo-adjacent homology domain; BROMO, Bromo domain; PAS, Per-Arnt-Sim domain. Size of scheme is proportional to the actual sequence length.



Supplementary Figure S19. Notch-like proteins of *M. brevicollis* and *C. owczarzaki*
Domain architectures of human Notch 1 and *M. brevicollis* and *C. owczarzaki* Notch-like proteins are schematically drawn. ANK; Ankyrin repeats; EGF, EGF-like domain; NL, Notch/Lin-12 repeat; NOD, Notch protein domain; NODP, NODP domain; SP, signal peptide; TM, transmembrane. Sizes of diagrams are proportional to the actual sequence lengths.



Supplementary Figure S20. Delta-like protein of *C. owczarzaki*
Domain architectures of human Delta 1 and one of the *C. owczarzaki* Delta-like proteins are schematically drawn. The *C. owczarzaki* protein was manually predicted from the nucleotide region 510391-515839 of the supercontig 12. Calx-beta, domain in Na-Ca exchangers and integrin subunit beta4; DSL, Delta/Serrate/lag-2 domain; EGF, EGF-like domain; MNNL, N-terminus of Notch ligand; SP, signal peptide; sushi, sushi domain; TM, transmembrane; TyrK, protein tyrosine kinase catalytic domain. Sizes of diagrams are proportional to the actual sequence lengths.



Supplementary Figure S21. Conservation of genes involved in Notch signaling
a, *C. owczarzaki* has Notch-like transmembrane proteins and Delta-like transmembrane proteins but their structures are still considerably different from the metazoan Notch and Delta proteins. b, *M. brevicollis* Notch-like protein had also been described[9,57]. c, ADAM10 and ADAM17 appears to have diverged in the early holozoan lineage and *C. owczarzaki* has lost the latter.

CAOG number	group	family	closest metazoan genes	*Mbre*
CAOG_04761	AGC	Akt*	AKT1/2/3	●
Sc9 1239598-1242669 (CAOG_05758)	AGC	Akt*	AKT1/2/3	●
CAOG_00316	AGC	DMPK*	Rock / Cdc42-binding protein kinase	●
CAOG_02802	AGC	DMPK*	DMPK-I/II / Genghis kinase	●
CAOG_03654	AGC	DMPK*		
CAOG_04705	AGC	DMPK*		
CAOG_06515	AGC	DMPK*		
CAOG_07198	AGC	DMPK*		
CAOG_06908	AGC	GRK*	GPCR / β-adrenergic receptor kinases	●
CAOG_00218	AGC	Mast*	Mast	●
CAOG_00305	AGC	Mast*	greatwall	●
CAOG_00905	AGC	Mast*	greatwall	●
CAOG_01870	AGC	Mast*		
CAOG_00619	AGC	NDR*	LATS (large tumor suppressor)	
CAOG_05338	AGC	NDR*	fungi / slime mold DBF2/20	
CAOG_07230	AGC	NDR*	NDR1/2	●
CAOG_05426	AGC	PDK1*	PkB / PDPK2	●
CAOG_01554	AGC	PKA*	PRKX / PRKY	
CAOG_02108	AGC	PKA*		
CAOG_02813	AGC	PKA*		
CAOG_02830	AGC	PKA*	PKA catalytic subunut α/β/γ	●
CAOG_06393	AGC	PKA*	PKA catalytic subunut α/β/γ	●
CAOG_00955	AGC	PKC*	PKC ι/ζ	
CAOG_02565	AGC	PKC*	PKC η/ε	●
CAOG_08272	AGC	PKN	PKN1/2	
CAOG_02592	AGC	RSK*	S6 kinase 90kDa / Jil-1	●
CAOG_05877	AGC	RSK*	S6 kinase 70kDa α/β	●
CAOG_00420	AGC	RSK*	Sgk	●
CAOG_01511	AGC	YANK	YANK1/2/3	
CAOG_05563	AGC	YANK	YANK1/2/3	
CAOG_01911	AGC	unclassified	fungi YPK1 / YKR2	●
CAOG_03553	AGC	unclassified	fungi unclassified kinase	
CAOG_05660	AGC	unclassified	fungi YPK1 / YKR2	●
CAOG_07765	AGC	unclassified		
CAOG_03763	CaMK	CaMK1*	CaMK1α/β/γ/δ	
CAOG_07915	CaMK	CaMK2*	CaMKIIα/β/γ/δ	●
CAOG_00377	CaMK	CaMKL*		
CAOG_08549	CaMK	CaMKL*	MAK-V / Hunk	●
CAOG_02014	CaMK	CaMKL*	LKB1	●
CAOG_02702	CaMK	CaMKL*		
CAOG_02898	CaMK	CaMKL*	PASK	●
CAOG_03076	CaMK	CaMKL*	MARK1/2/3/4	●
Sc4 1536259-1540513 (CAOG_08683)	CaMK	CaMKL*		
CAOG_04330	CaMK	CaMKL*	NUAK1/2	
CAOG_05898	CaMK	CaMKL*	AMPKα1/2	●
CAOG_06600	CaMK	CaMKL*		
CAOG_06772	CaMK	CaMKL*	SNF-related kinase	●
CAOG_09042	CaMK	CaMKL*	SADA/B	
CAOG_08204	CaMK	CaMKL*	NUAK1/2	

Supplementary Figure S22 continued

CAOG number	group	family	closest metazoan genes	*Mbre*
CAOG_05773	CaMK	MAPKAPK*	MNK1/2	●
CAOG_08994	CaMK	MAPKAPK*	MAPKAP kinase2/3	●
CAOG_05910	CaMK	PHK	PHKγ1/2	
CAOG_00433	CaMK	PIM	PIM1/2/3	
CAOG_02400	CaMK	PKD	PKD1/2/3	
CAOG_02958	CaMK	PKD	PKD1/2/3	
CAOG_03061	CaMK	RAD53*	Chk2	●
CAOG_01948	CaMK	Trb	TRB (tribbles)	
CAOG_02197	CaMK	unclassified		
CAOG_02304	CaMK	unclassified		●
CAOG_03105	CaMK	unclassified	(Similar domain architecture to PASK)	
CAOG_03852	CaMK	unclassified		
CAOG_04000	CaMK	unclassified		
CAOG_04185	CaMK	unclassified		
CAOG_04619	CaMK	unclassified		
CAOG_04974	CaMK	unclassified		●
CAOG_05134	CaMK	unclassified		
CAOG_05447	CaMK	unclassified		
CAOG_08985	CaMK	unclassified		
CAOG_07003	CaMK	unclassified		
CAOG_07193	CaMK	unclassified		
CAOG_07315	CaMK	unclassified		
CAOG_02768	CK1	CK1*	Casein kinase 1γ	●
CAOG_03955	CK1	CK1*		
CAOG_07938	CK1	CK1*		
CAOG_00828	CK1	unclassified		
CAOG_00023	CMGC	CDK*	CDK5	●
CAOG_08444	CMGC	CDK*	PITSLRE	●
CAOG_01914	CMGC	CDK*	CDK7	●
CAOG_01997	CMGC	CDK*	PCTAIRE / PFTAIRE	●
CAOG_02335	CMGC	CDK*	CRK7 / CHED	●
CAOG_07047	CMGC	CDK*	PCTAIRE / PFTAIRE	●
CAOG_07268	CMGC	CDK*		
CAOG_07450	CMGC	CDK*	CDK9	●
CAOG_07905	CMGC	CDK*	CDC2	●
CAOG_06855	CMGC	CK2*	casein kinase 2 α1/2	●
CAOG_02016	CMGC	CMGC-I / CDK*	CDK8 / CDK8-like	●
CAOG_00863	CMGC	DYRK*	DYRK1	●
CAOG_03006	CMGC	DYRK*	amoebozoan, fungi and plants Yak	
CAOG_03480	CMGC	DYRK*	PRP4	●
CAOG_04106	CMGC	DYRK*	plants and protists Dyrk-related	
Sc1 2658011-2659256 (CAOG_00834)	CMGC	GSK*	GSK3	
CAOG_02578	CMGC	MAPK / Erk*	slime mold erkA	
CAOG_04095	CMGC	MAPK / Erk*	p38α/β/γ/ζ	●
CAOG_05850	CMGC	MAPK / Erk*	Erk5	●
CAOG_08222	CMGC	MAPK / Erk*	Erk1/2	●
CAOG_03488	CMGC	SRPK*		
CAOG_05609	CMGC	unclassified	basal metazoan uncharacterized CMGC group genes	●
Sc1 289584-283955 (CAOG_00078)	STE	STE7*		
CAOG_00295	STE	STE7*	MEK5	●

Supplementary Figure S22 continued

CAOG number	group	family	closest metazoan genes	*Mbre*
CAOG_01638	STE	STE7*	MEK1/2	●
CAOG_05490	STE	STE7*	MEK3/6/4/7	●
CAOG_03065	STE	STE11*		
CAOG_03579	STE	STE11*	MEKK2/3	●
Sc5 1441692-1434338 (CAOG_03874)	STE	STE11*	MEKK4	
CAOG_04640	STE	STE11*	Tpl-2	
CAOG_06694	STE	STE11*	MEKK1	
CAOG_08033	STE	STE11*	fungi and slime mold septation (sep) -related	
CAOG_00112	STE	STE20*	TAO1/2/3	●
CAOG_00558	STE	STE20*	KHS1/2 / HPK1	●
CAOG_00822	STE	STE20*	slime mold PakE/F/G/H	
CAOG_00846	STE	STE20*	LYK (STLK)	●
CAOG_00922	STE	STE20*	MASK / STK25 / STK24 / Mst3	●
CAOG_01322	STE	STE20*	MASK / STK25 / STK24 / Mst3	●
CAOG_01226	STE	STE20*	TAO	●
CAOG_01320	STE	STE20*	PAK1/2/3	●
CAOG_06041	STE	STE20*	PAK1/2/3	●
CAOG_01932	STE	STE20*	MST1/2	●
CAOG_02628	STE	STE20*		
CAOG_04691	STE	STE20*		
CAOG_04997	STE	STE20*	SLK / LOK	
CAOG_07614	STE	STE20*	OSR1 / SPAK	
CAOG_08084	STE	STE20*		
CAOG_00099	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_00335	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_01039	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_01146	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_01363	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_01599	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_01820	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_01933	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_01994	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_01995	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_02213	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_08617	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_02560	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_02782	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_03008	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_03051	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_03058	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_03145	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
Sc4 1436556-1432639 (CAOG_08675)	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_03272	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_03490	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_03611	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_03670	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_03786	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
Sc5 1614159-1611980 (CAOG_03923)	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_03948	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_04160	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_04351	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	

Supplementary Figure S22 continued

CAOG number	group	family	closest metazoan genes	*Mbre*
CAOG_04555	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_04596	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_04911	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_04928	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05144	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05255	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05294	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_08893	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05571	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05592	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05649	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05665	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05677	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05756	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05762	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05771	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05780	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05783	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05812	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05960	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_06176	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_06219	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_06294	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_06330	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_06468	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_06497	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_06567	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07124	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07148	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07150	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07151	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07155	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_09054	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07233	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07237	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07350	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07517	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07542	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07591	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07748	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07749	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07803	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07810	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_09128	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07914	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07926	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_07934	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_09154	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_08213	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_08263	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	

Supplementary Figure S22 continued

CAOG number	group	family	closest metazoan genes	*Mbre*
CAOG_08675	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_08838	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_08896	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_09061	TKL	CoLCSTK / IRAK	IRAK1/2/3/4	
CAOG_05847	TKL	LISK*	LIMK1/2 / TESK1/2	●
CAOG_01131	TKL	MLK*	MLK1/2/3/4	●
CAOG_01366	TKL	MLK*	ILK (integrin-linked kinase)	
CAOG_01546	TKL	MLK*	LZK / ZPK	●
CAOG_02126	TKL	MLK*	TNNI3K	●
CAOG_04570	TKL	MLK*	ZAK	
CAOG_05658	TKL	RAF	A/B/C-raf	
CAOG_08997	TKL	RAF	A/B/C-raf	
CAOG_01057	TKL	unclassified		
CAOG_03558	TKL	unclassified		
CAOG_04428	TKL	unclassified		
CAOG_04937	TKL	unclassified	choanoflagellates' uncharacterized TKL group genes	●
CAOG_05099	TKL	unclassified		
CAOG_07634	TKL	unclassified		
CAOG_08335	TKL	unclassified		
CAOG_06708	Protein Tyrosine Kinase	Abl*	Abl / Arg	●
CAOG_00534	Protein Tyrosine Kinase	Csk*	Csk	●
CAOG_02460	Protein Tyrosine Kinase	Csk*	Csk	●
CAOG_06750	Protein Tyrosine Kinase	Fak	FAK / PYK2	
CAOG_02182	Protein Tyrosine Kinase	Src*	Src / Yes / Yrk / Fyn / Fgr / Lyn / Hck / Lck / Blk	●
CAOG_06360	Protein Tyrosine Kinase	Src*	Src / Yes / Yrk / Fyn / Fgr / Lyn / Hck / Lck / Blk	●
CAOG_03369	Protein Tyrosine Kinase	Tec*	TEC / TXK / BTK / BMX / ITK	●
Sc1 112601-109934 (CAOG_00027)	Protein Tyrosine Kinase	CoLRYK		
CAOG_00226	Protein Tyrosine Kinase	CoLRYK		
CAOG_00604	Protein Tyrosine Kinase	CoLRYK		
CAOG_01090	Protein Tyrosine Kinase	CoLRYK		
CAOG_08470	Protein Tyrosine Kinase	CoLRYK		
CAOG_08481	Protein Tyrosine Kinase	CoLRYK		
CAOG_01170	Protein Tyrosine Kinase	CoLRYK		
CAOG_01470	Protein Tyrosine Kinase	CoLRYK		
CAOG_08531	Protein Tyrosine Kinase	CoLRYK		
CAOG_01676	Protein Tyrosine Kinase	CoLRYK		
CAOG_01840	Protein Tyrosine Kinase	CoLRYK		
CAOG_01970	Protein Tyrosine Kinase	CoLRYK		
CAOG_02086	Protein Tyrosine Kinase	CoLRYK		
CAOG_02360	Protein Tyrosine Kinase	CoLRYK		
CAOG_02447	Protein Tyrosine Kinase	CoLRYK		
CAOG_02494	Protein Tyrosine Kinase	CoLRYK		
CAOG_02517	Protein Tyrosine Kinase	CoLRYK		
CAOG_02623	Protein Tyrosine Kinase	CoLRYK		
CAOG_08668	Protein Tyrosine Kinase	CoLRYK		
CAOG_08701	Protein Tyrosine Kinase	CoLRYK		
CAOG_03572	Protein Tyrosine Kinase	CoLRYK		
CAOG_03667	Protein Tyrosine Kinase	CoLRYK		
CAOG_03717	Protein Tyrosine Kinase	CoLRYK		

Supplementary Figure S22 continued

CAOG number	group	family	closest metazoan genes	*Mbre*
CAOG_03740	Protein Tyrosine Kinase	CoLRYK		
CAOG_03855	Protein Tyrosine Kinase	CoLRYK		
CAOG_04365	Protein Tyrosine Kinase	CoLRYK		
CAOG_04368	Protein Tyrosine Kinase	CoLRYK		
CAOG_04404	Protein Tyrosine Kinase	CoLRYK		
CAOG_04506	Protein Tyrosine Kinase	CoLRYK		
CAOG_04557	Protein Tyrosine Kinase	CoLRYK		
CAOG_04598	Protein Tyrosine Kinase	CoLRYK		
CAOG_04757	Protein Tyrosine Kinase	CoLRYK		
CAOG_04811	Protein Tyrosine Kinase	CoLRYK		
CAOG_04919	Protein Tyrosine Kinase	CoLRYK		
CAOG_04955	Protein Tyrosine Kinase	CoLRYK		
CAOG_05077	Protein Tyrosine Kinase	CoLRYK		
CAOG_05156	Protein Tyrosine Kinase	CoLRYK		
CAOG_05181	Protein Tyrosine Kinase	CoLRYK		
CAOG_05278	Protein Tyrosine Kinase	CoLRYK		
CAOG_08875	Protein Tyrosine Kinase	CoLRYK		
CAOG_05321	Protein Tyrosine Kinase	CoLRYK		
CAOG_05325	Protein Tyrosine Kinase	CoLRYK		
CAOG_05356	Protein Tyrosine Kinase	CoLRYK		
CAOG_05394	Protein Tyrosine Kinase	CoLRYK		
CAOG_05455	Protein Tyrosine Kinase	CoLRYK		
CAOG_05532	Protein Tyrosine Kinase	CoLRYK		
CAOG_05636	Protein Tyrosine Kinase	CoLRYK		
CAOG_05691	Protein Tyrosine Kinase	CoLRYK		
CAOG_05733	Protein Tyrosine Kinase	CoLRYK		
CAOG_05761	Protein Tyrosine Kinase	CoLRYK		
CAOG_05796	Protein Tyrosine Kinase	CoLRYK		
CAOG_05854	Protein Tyrosine Kinase	CoLRYK		
CAOG_05958	Protein Tyrosine Kinase	CoLRYK		
CAOG_06007	Protein Tyrosine Kinase	CoLRYK		
CAOG_06160	Protein Tyrosine Kinase	CoLRYK		
CAOG_06234	Protein Tyrosine Kinase	CoLRYK		
CAOG_06280	Protein Tyrosine Kinase	CoLRYK		
CAOG_06282	Protein Tyrosine Kinase	CoLRYK		
CAOG_06310	Protein Tyrosine Kinase	CoLRYK		
CAOG_06366	Protein Tyrosine Kinase	CoLRYK		
CAOG_06420	Protein Tyrosine Kinase	CoLRYK		
CAOG_06439	Protein Tyrosine Kinase	CoLRYK		
CAOG_06570	Protein Tyrosine Kinase	CoLRYK		
CAOG_06582	Protein Tyrosine Kinase	CoLRYK		
CAOG_06608	Protein Tyrosine Kinase	CoLRYK		
CAOG_06673	Protein Tyrosine Kinase	CoLRYK		
CAOG_06751	Protein Tyrosine Kinase	CoLRYK		
CAOG_06841	Protein Tyrosine Kinase	CoLRYK		
CAOG_06945	Protein Tyrosine Kinase	CoLRYK		
CAOG_06958	Protein Tyrosine Kinase	CoLRYK		
CAOG_06961	Protein Tyrosine Kinase	CoLRYK		
CAOG_06964	Protein Tyrosine Kinase	CoLRYK		

Supplementary Figure S22 continued

CAOG number	group	family	closest metazoan genes	*Mbre*
CAOG_06991	Protein Tyrosine Kinase	CoLRYK		
CAOG_09045	Protein Tyrosine Kinase	CoLRYK		
CAOG_09046	Protein Tyrosine Kinase	CoLRYK		
CAOG_07211	Protein Tyrosine Kinase	CoLRYK		
CAOG_07406	Protein Tyrosine Kinase	CoLRYK		
CAOG_07702	Protein Tyrosine Kinase	CoLRYK		
CAOG_09123	Protein Tyrosine Kinase	CoLRYK		
CAOG_07752	Protein Tyrosine Kinase	CoLRYK		
CAOG_07757	Protein Tyrosine Kinase	CoLRYK		
CAOG_09125	Protein Tyrosine Kinase	CoLRYK		
CAOG_08089	Protein Tyrosine Kinase	CoLRYK		
CAOG_08103	Protein Tyrosine Kinase	CoLRYK		
Sc17 595782-591943 (CAOG_09167)	Protein Tyrosine Kinase	CoLRYK		
CAOG_08168	Protein Tyrosine Kinase	CoLRYK		
CAOG_09181	Protein Tyrosine Kinase	CoLRYK		
CAOG_08227	Protein Tyrosine Kinase	CoLRYK		
CAOG_08238	Protein Tyrosine Kinase	CoLRYK		
CAOG_08239	Protein Tyrosine Kinase	CoLRYK		
CAOG_09189	Protein Tyrosine Kinase	CoLRYK		
CAOG_08241	Protein Tyrosine Kinase	CoLRYK		
CAOG_08274	Protein Tyrosine Kinase	CoLRYK		
CAOG_08297	Protein Tyrosine Kinase	CoLRYK		
Sc41 2908-6928 (CAOG_08328)	Protein Tyrosine Kinase	CoLRYK		
CAOG_08354	Protein Tyrosine Kinase	CoLRYK		
CAOG_08364	Protein Tyrosine Kinase	CoLRYK		
CAOG_08366	Protein Tyrosine Kinase	CoLRYK		
CAOG_02075	Protein Tyrosine Kinase	unclassified		
CAOG_00535	Other	Aur*	Aurora	●
CAOG_01569	Other	BUB	BUB	
CAOG_06610	Other	CaMKK*	CaMKKα/β	●
CAOG_09025	Other	CaMKK*		
CAOG_09033	Other	Haspin	haspin	
CAOG_01915	Other	IRE*		
CAOG_04163	Other	IRE*	IRE1/2	●
CAOG_00354	Other	NAK*	GAK	
CAOG_00763	Other	NAK*	BIKE	
CAOG_08028	Other	NAK*	MPSK1	
CAOG_00759	Other	NEK*		
CAOG_02107	Other	NEK*		●
CAOG_06703	Other	NRBP*	NRBP1/2	●
CAOG_03741	Other	PEK*	GCN2	●
CAOG_03792	Other	PEK*		
CAOG_07848	Other	PLK*	polo-like kinase 1/2/3/4/5	●
CAOG_02941	Other	SCY1*	SCY1-like1	●
CAOG_08871	Other	SCY1*	SCY1-like3	●
CAOG_08172	Other	SgK196	SgK196	
CAOG_02145	Other	Slob*	modulator of Na,K-ATPase / Slob	●
CAOG_02207	Other	TLK*	Tlk(tousled-like kinase)1 / Tlk2	●
CAOG_04389	Other	TTK*	Ttk	●

Supplementary Figure S22 continued

CAOG number	group	family	closest metazoan genes	*Mbre*
CAOG_04769	Other	ULK*	ULK1/2	•
CAOG_04326	Other	VPS15*	PI3K regulatory subunit	•
CAOG_08396	Other	Wee*	Myt	
CAOG_09184	Other	Wee*	Wee1/1B	•
Sc6 437808-444525 (CAOG_04111)	Other	WNK*	WNK1/2/3/4	•
CAOG_00311	Other	unclassified		
CAOG_00549	Other	unclassified		•
CAOG_01202	Other	unclassified		
CAOG_01845	Other	unclassified		
CAOG_01895	Other	unclassified		
CAOG_01986	Other	unclassified		
CAOG_02000	Other	unclassified		
CAOG_08582	Other	unclassified		
CAOG_02457	Other	unclassified		
CAOG_02899	Other	unclassified		
CAOG_03445	Other	unclassified		
CAOG_03665	Other	unclassified		
CAOG_03849	Other	unclassified		
CAOG_03868	Other	unclassified		
CAOG_03922	Other	unclassified		
CAOG_04299	Other	unclassified		
CAOG_04467	Other	unclassified		
Sc7 514955-512626 (CAOG_04653)	Other	unclassified		
CAOG_04961	Other	unclassified	PAB-dependent poly(A)-specific ribonuclease subunit	•
CAOG_08864	Other	unclassified		
CAOG_05331	Other	unclassified		
CAOG_05528	Other	unclassified		
CAOG_06029	Other	unclassified		
CAOG_09003	Other	unclassified		
CAOG_07325	Other	unclassified		
CAOG_07343	Other	unclassified		
CAOG_07360	Other	unclassified		
CAOG_07552	Other	unclassified		
CAOG_07646	Other	unclassified		
CAOG_08035	Other	unclassified		
CAOG_08375	Other	unclassified		
CAOG_08563	Other	unclassified		

Supplementary Figure S22. *C. owczarzaki* protein kinases
The 384 PKs found in the genome of *C. owczarzaki* are classified into 8 groups (AGC, CaMK, CK1, CMGC, STE, TKL, Protein Tyrosine Kinase, and Other), which are further subdivided into families. Genes overlooked or mis-predicted by the automatic annotation pipeline were manually re-predicted and their coordinates in the supercontig (Sc), instead of the CAOG numbers, are shown. PK families present also in the *M. brevicollis* genome are labelled by asterisks. Next to the family classification, the closest metazoan genes (unless otherwise specified) on the basis of the kinase domain phylogeny are shown. Presence of putative *M. brevicollis* (Mbre) orthologs are shown by dots on the right.

Supplementary Figure S23. Structures of CoLCSTKs
Six representative architectures of CoLCSTK family genes are schematically shown. Names of the other CoLCSTKs having the identical architecture are shown in parentheses, except for the typical CoLCSTK class, where only the gene number is shown. KD, kinase domain; LRR, leucine rich repeat; MSP, major sperm protein domain; RAS, RAS small GTPase domain; SH2, Src homology 2 domain; Ulp1, Ulp1 protease family C-terminal catalytic domain; ZF, $Cys_3HisCys_4$ zinc finger.

MAPKKK

Family	Gene	METAZOA *Hs*	*Ta*	*Aq*	*Mb*	***Co***	*Sc*	*Dd*	*Cr*
STE11	MEKK1	●	●		●	●			
	MEKK2	●	●	●	●	●			
	MEKK3	●							
	MEKK4	●	●	●		●	●		
	ASK1	●							
	ASK2	●	●	●	●				
	MEKK15	●							
	Tpl2	●	●	●		●			
	STE11						●		●
	BCK1						●		
STE20	TAO1	●							
	TAO2	●	●	●	●	●			
	TAO3	●							
TKL	MLK1	●							
	MLK2	●							
	MLK3	●	●	●	●	●			
	MLK4	●							
	ZAK	●		●		●			
	LZK	●							
	DLK	●	●	●	●	●			
	A-RAF	●							
	B-RAF	●	●	●		●			
	C-RAF	●							
	TAK1	●	●	●					
Other	MOS	●	●						
	NIK	●							

Supplementary Figure S24 continued

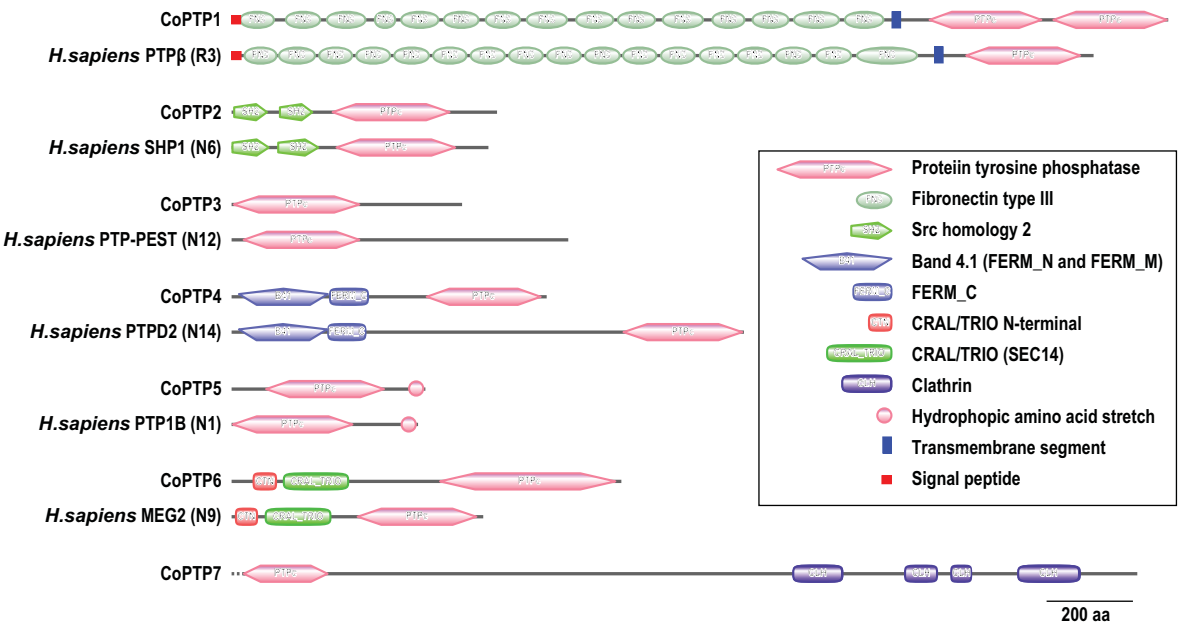
MAPKK

		METAZOA							
Family	Gene	*Hs*	*Ta*	*Aq*	*Mb*	***Co***	*Sc*	*Dd*	*Cr*
STE7	MEK1	●	●	●	●	●	●		
	MEK2	●							
	MKK3	●	●	●					
	MKK6	●				●			
	MKK4	●	●	●				●	●
	MKK7	●	●	●					
	MKK5	●		●	●	●			
	STE7						●		
	MKK1						●		

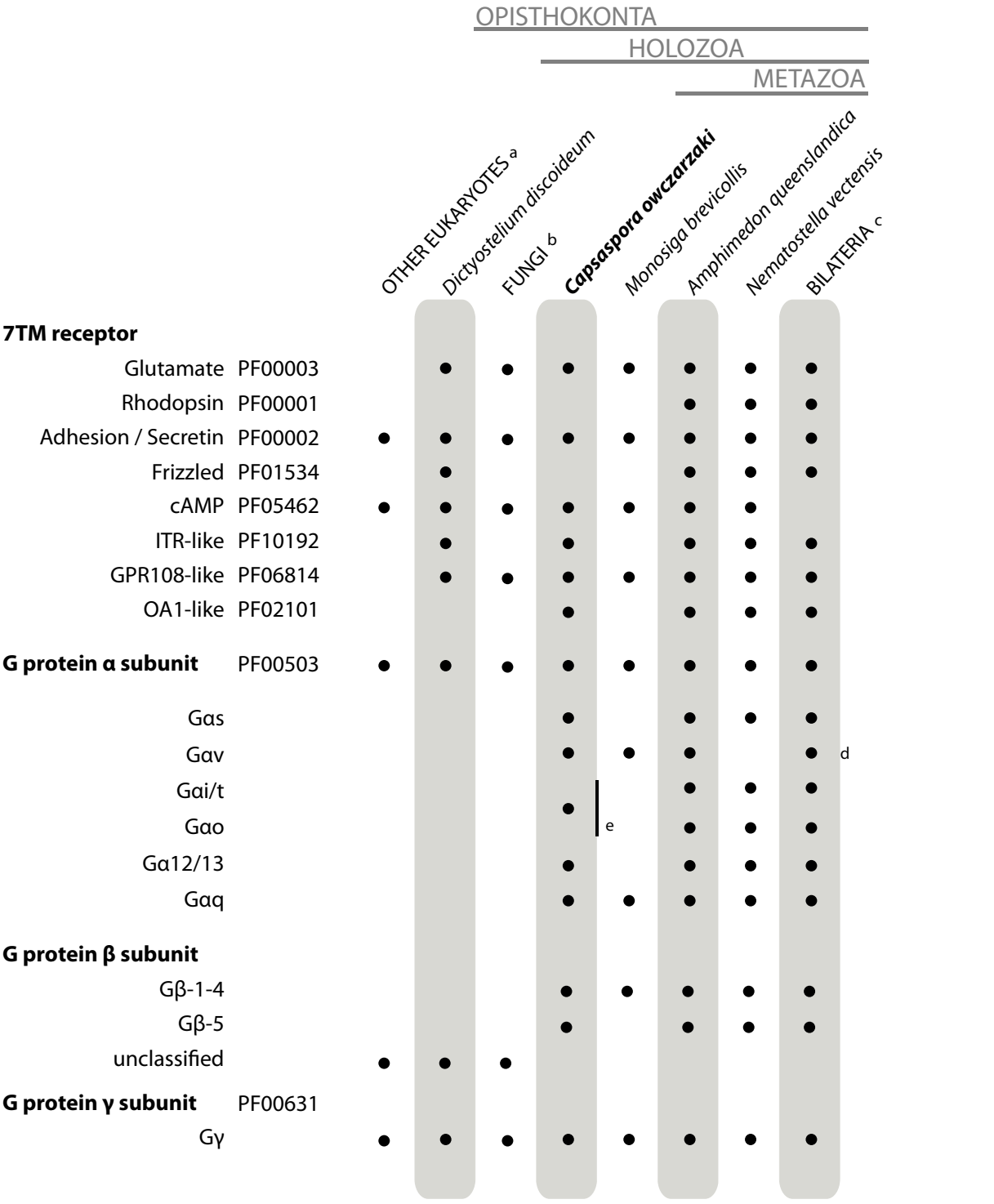
MAPK

		METAZOA							
Family	Gene	*Hs*	*Ta*	*Aq*	*Mb*	***Co***	*Sc*	*Dd*	*Cr*
MAPK	ERK1	●	●	●	●	●			
	ERK2	●					●		
	ERK3	●							
	ERK4	●							
	ERK5	●	●	●	●	●			
	JNK1	●							
	JNK2	●	●	●					
	JNK3	●						●	●
	p38α	●							
	p38β	●	●	●		●	●		
	p38γ	●							
	p38δ	●							
	NLK	●	●	●					
	SLT2	●					●		
	SMK1	●					●		
	ERK7	●	●	●				●	●

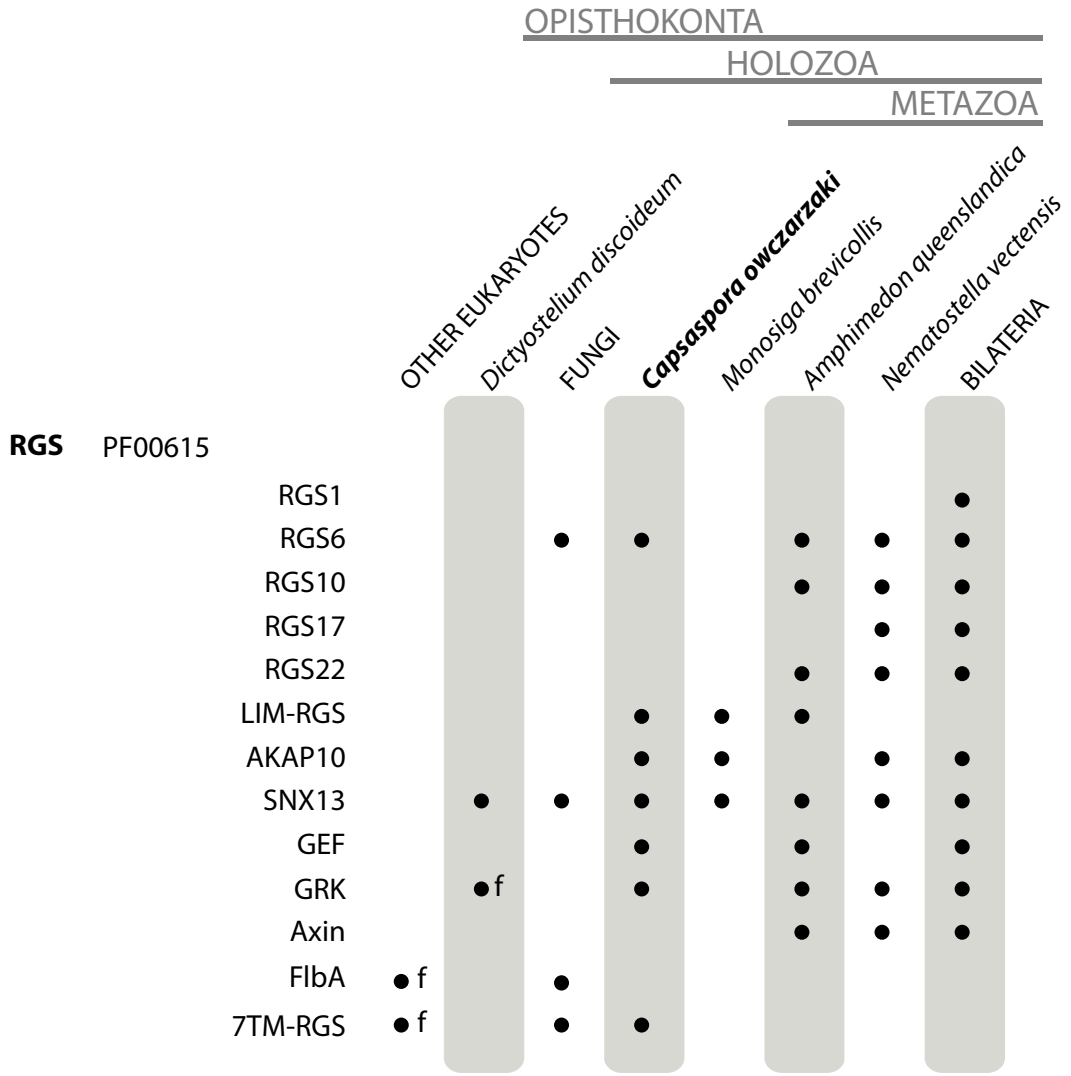
Supplementary Figure S24. Conservation of MAPK pathway genes
Eight whole genome sequences (Hs, *Homo sapiens*; Ta, *Trichoplax adhaerens*; Aq, *Amphimedon queenslandica*; Mb, *Monosiga brevicollis*; Co, *Capsaspora owczarzaki*; Sc, *Saccharomyces cerevisiae*; Dd, *Dictyostelium discoideum*; Cr, *Chlamydomonas reinhardtii*) were searched for the MAPK pathway components in three different classes, MAPK, MAPKK and MAPKKK[58-60]. Genes were classified into families according to http://www.kinase.com. Dots represent the presence of genes. A vertical line indicates a lineage-specific duplication or unclear orthology due to the weak phylogenetic signal.



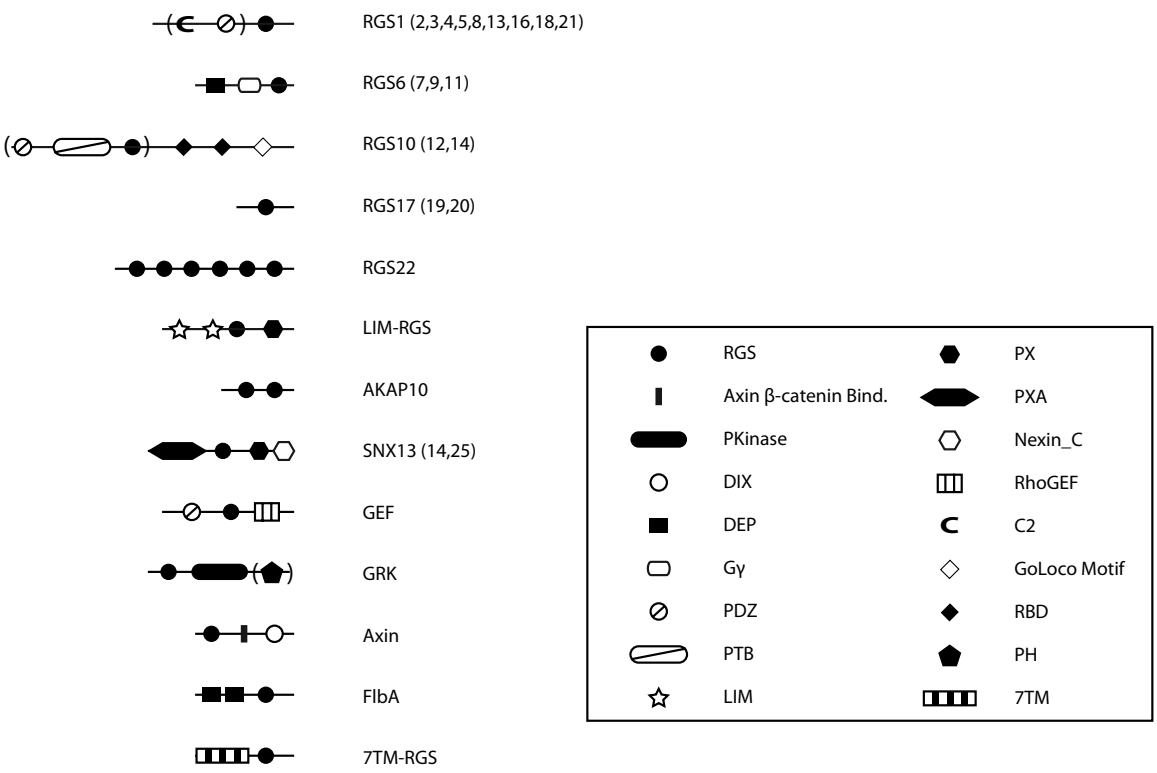
Supplementary Figure S25. *C. owczarzaki* PTPs
Domain architectures of the seven *C. owczarzaki* PTPs are schematically shown together with their putative human homologs. The N-terminal sequence of CoPTP7 is unknown due probably to a sequence gap or an assembly problem (dotted line). Amino acid sequences were manually predicted from genomic supercontigs in the following regions: CoPTP1, supercontig1 918228-911321; CoPTP2, supercontig13 292707- 295098; CoPTP3, supercontig2 979032-981976; CoPTP4, supercontig16 516981-519527; CoPTP5, supercontig2 2241603-2239462; CoPTP6, supercontig13 227923-224188; CoPTP7, supercontig7 511782-505305. Sizes of diagrams are proportional to the actual sequence lengths.



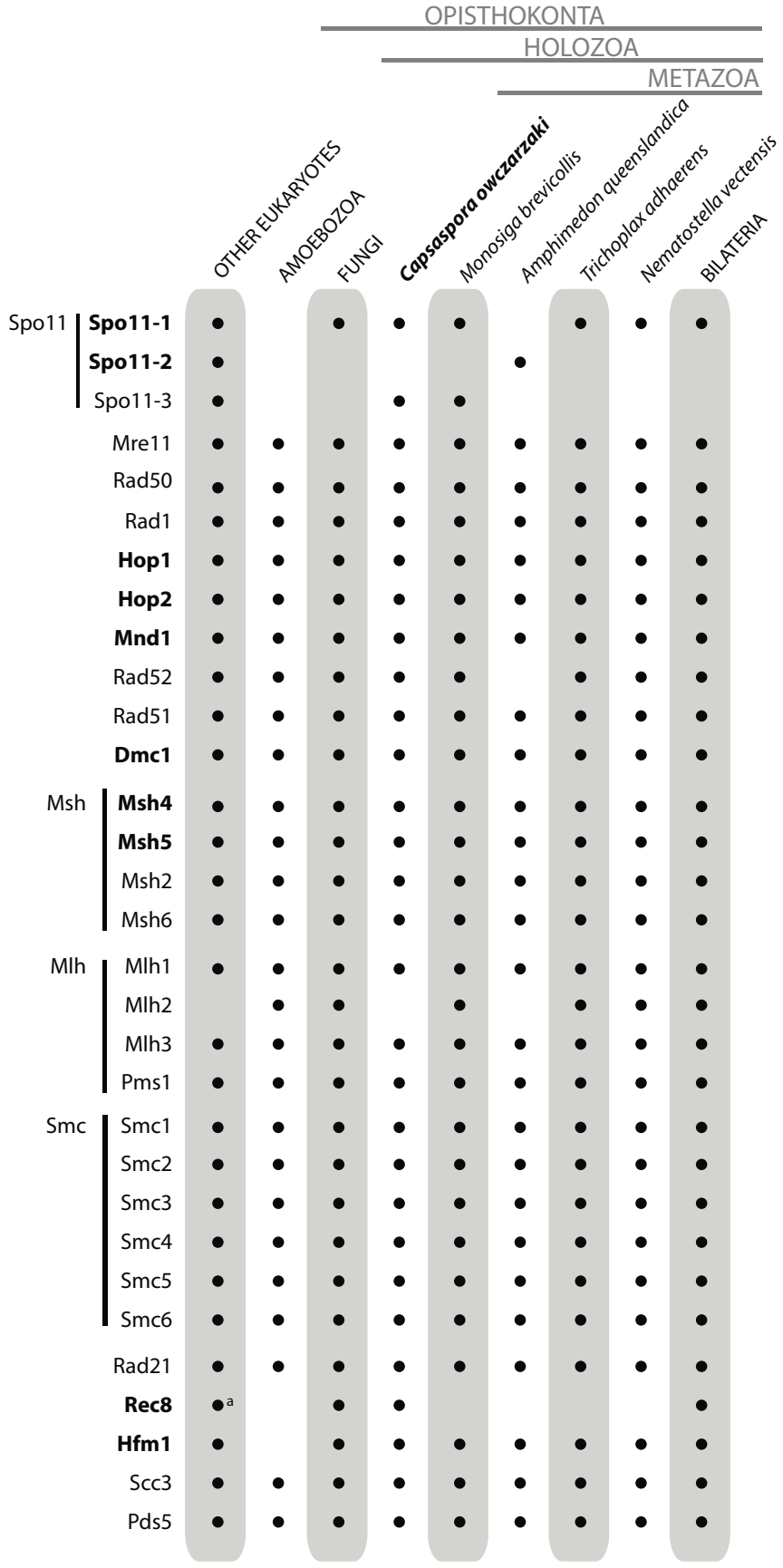
Supplementary Figure S26 continued



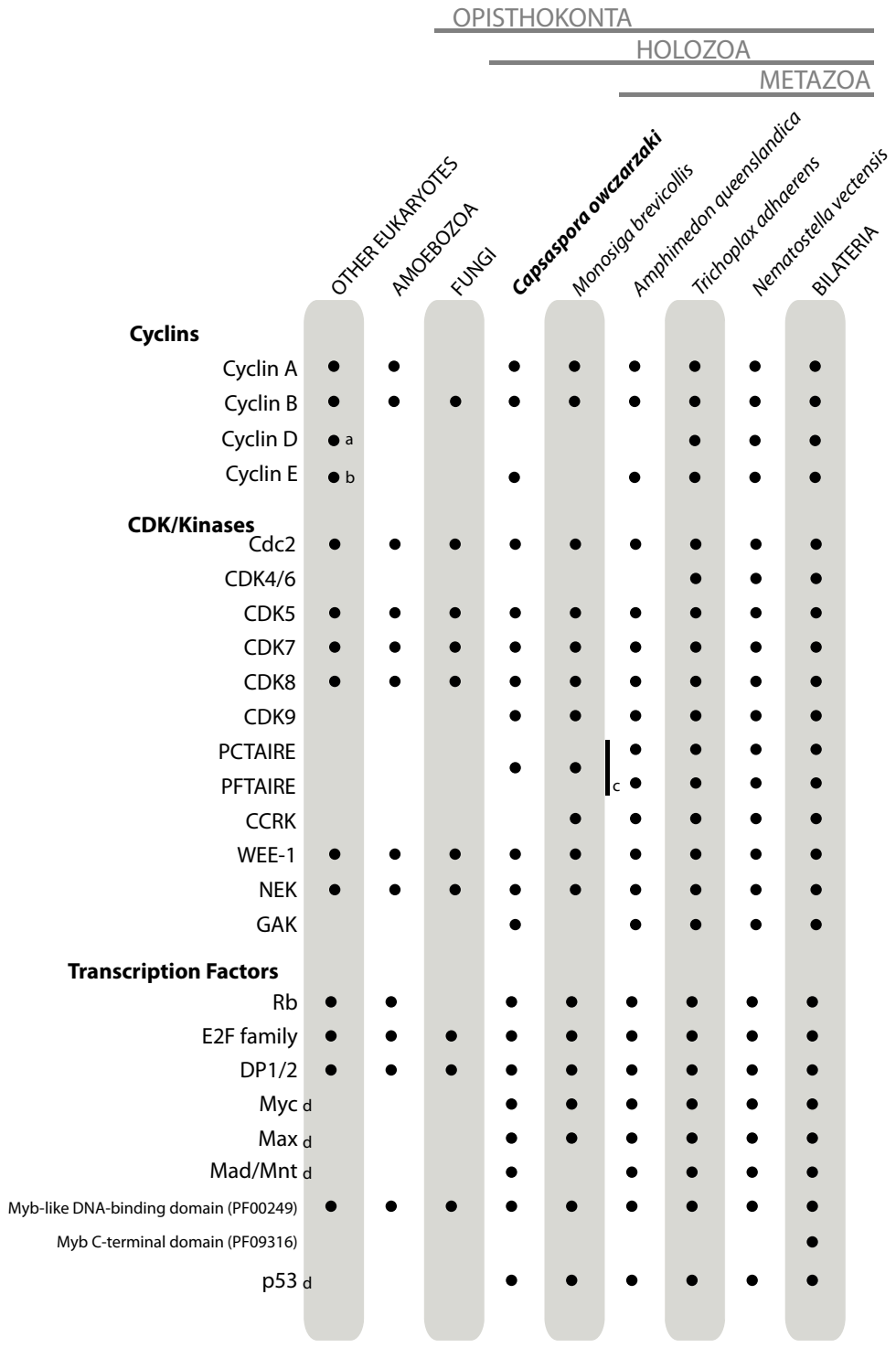
Supplementary Figure S26. Conservation of genes involved in 7TM receptors and their signaling
Presence of a gene is indicated by a dot. Pfam statistical models used for HMMER search are shown after family names when available. AKAP, A-kinase anchor protein; SNX, sorting nexin; 7TM-RGS, undescribed RGS family with seven transmembrane segments. a, the genomes of the amoeboflagellate *Naegleria gruberi* and some stramenopiles genomes were analyzed; b, The genomes of *Neurospora crassa*, *Ustilago* maydis, and *Schizosaccharomyces pombe*, were analyzed c, *H. sapiens* and *D. melanogaster* genomes were analyzed; d, although fish, basal chordates, and some arthropods have it, tetrapods and many other invertebrate species seem to have lost it[61]; e, Gαi/t and Gαo seem to have diverged in the lineage leading to choanoflagellates and metazoans after their divergence from filastereans; f, proteins with same domain architecture present in amoebozoans (GRK), *N. gruberi* (FlbA), and green plants, *Ectocarpus siliculosus*, and *Naegleria gruberi* (7TM-RGS) though not confidently classified to families by the sequence homology of RGS domain.



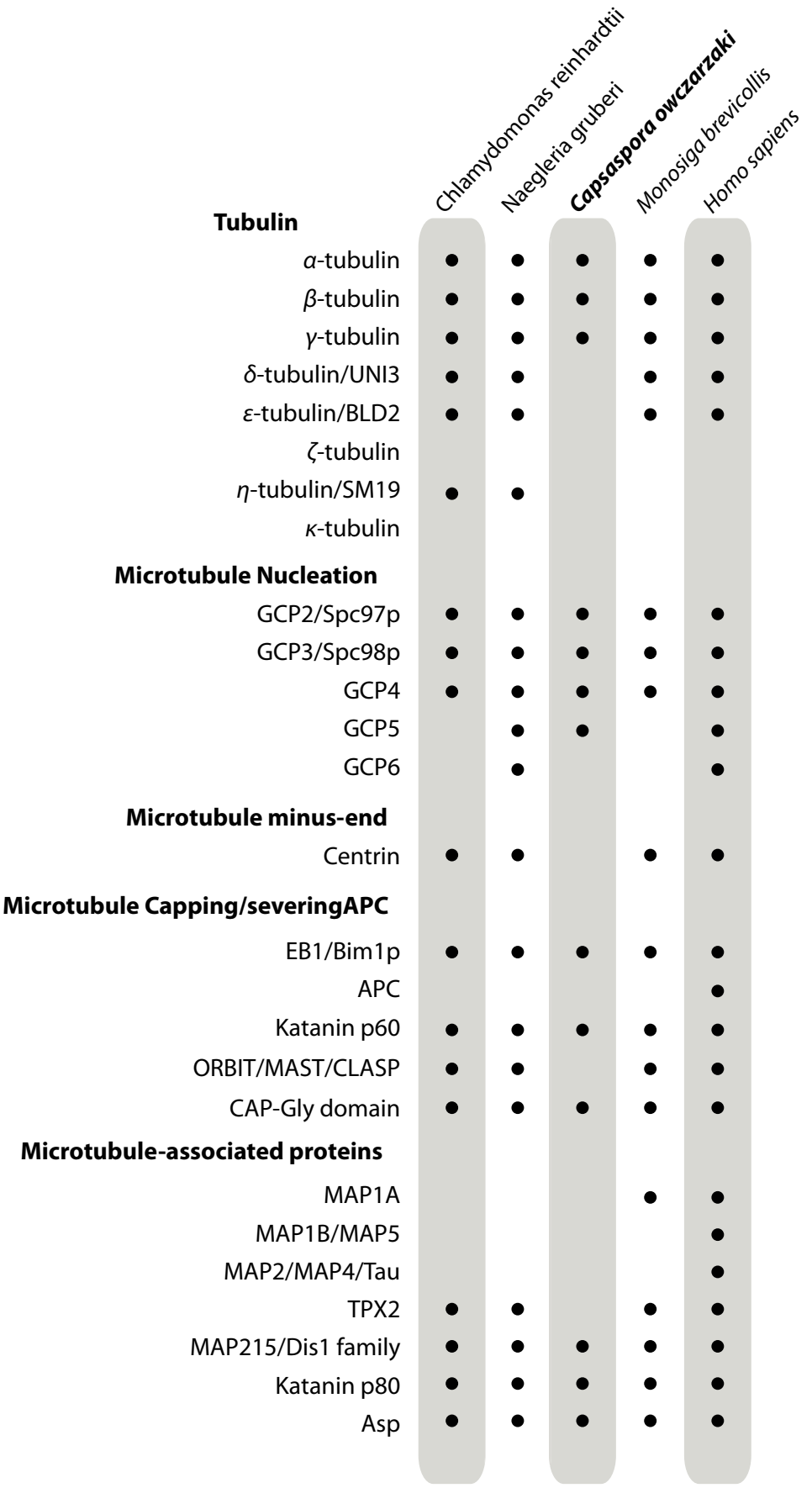
Supplementary Figure S27. Protein domain architectures of RGS families
Schematic drawings of domain architectures are shown for 13 RGS families that the *C. owczarzaki* genome encodes. Parentheses indicate either that domains are absent or have been added in some members. Pfam or SMART domain names of the schemes are shown in inset. 7TM, seven transmembrane segments; AKAP, A-kinase anchor protein; SNX, sorting nexin.



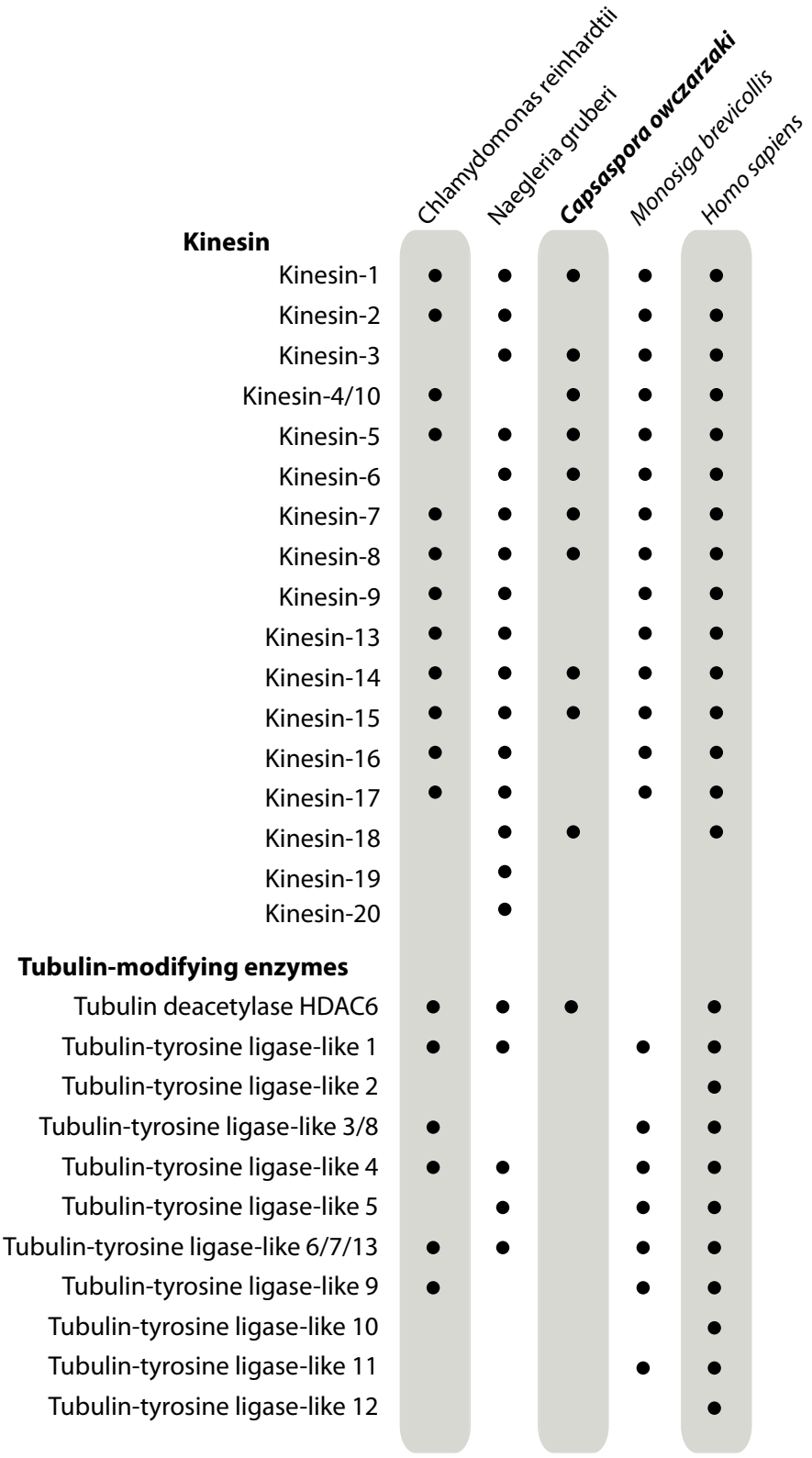
Supplementary Figure S28. Survey of meiotic genes in *C. owczarzaki* and other eukaryotes
Dots indicate presence of orthologs. Genes in bold letters are specifically involved in meiosis. a, present only in green plants.



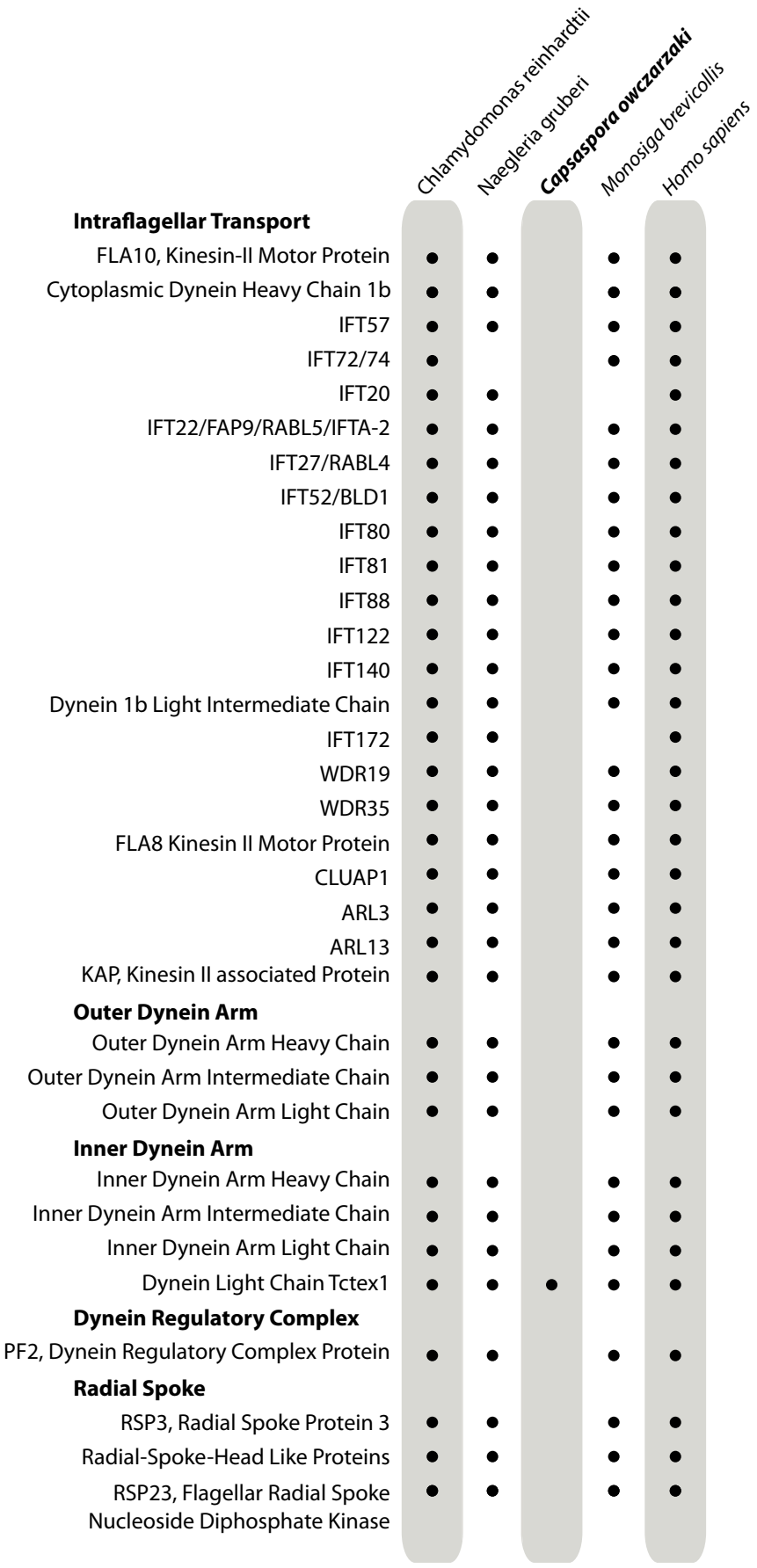
Supplementary Figure S29. Survey of cell cycle regulators
a, green plants have genes named cyclin D though their orthology to metazoan genes is not strongly supported by our phylogenetic analyses; b, the protists *Paramecium* seem to have cyclin E though phylogenetic signal is weak; c, PCTAIRE and PFTAIRE are likely to have diverged in the metazoan lineage after the separation from choanoflagellates; d, data taken from a previous publication[17].



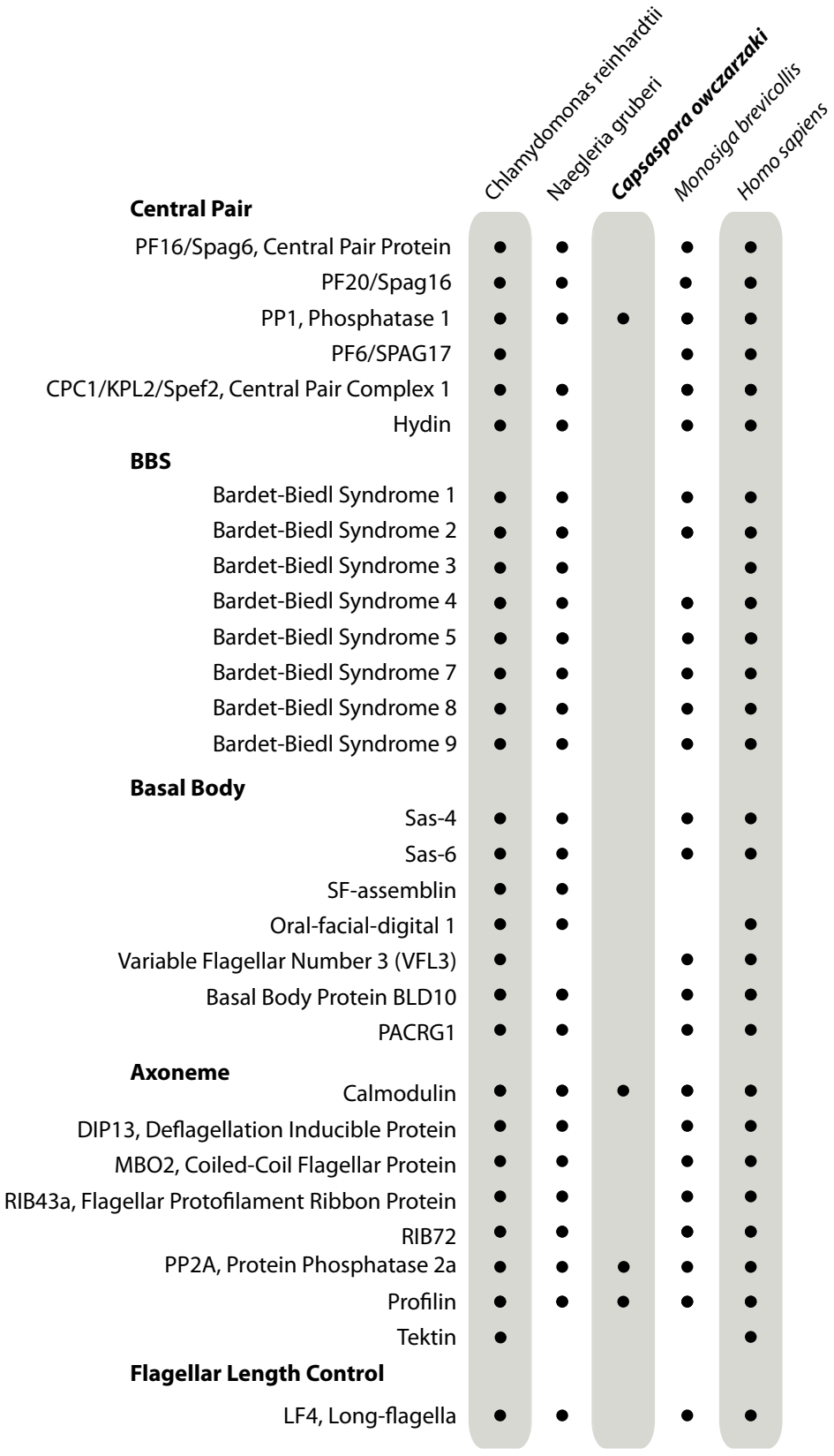
Supplementary Figure S30 continued



Supplementary Figure S30 continued



Supplementary Figure S30 continued

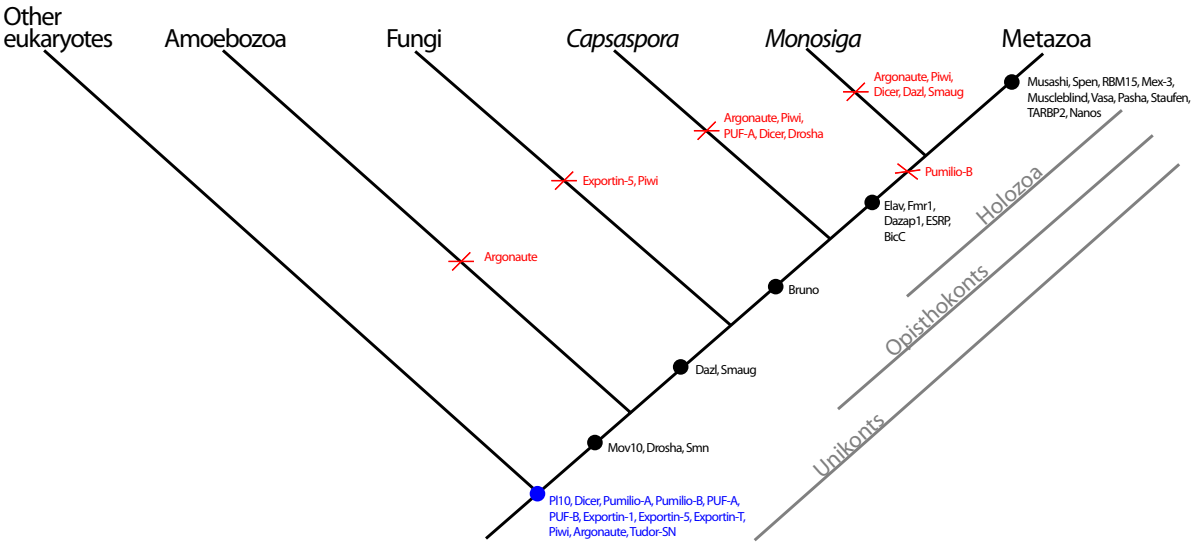


Supplementary Figure S30. Survey of flagellum genes
Four genomes of differently flagellated eukaryotes (*C. reinhardtii*, *N. gruberi*, *M. brevicollis*, and *H. sapiens*) were searched for flagellar apparatus genes and compared with that of *C. owczarzaki*.

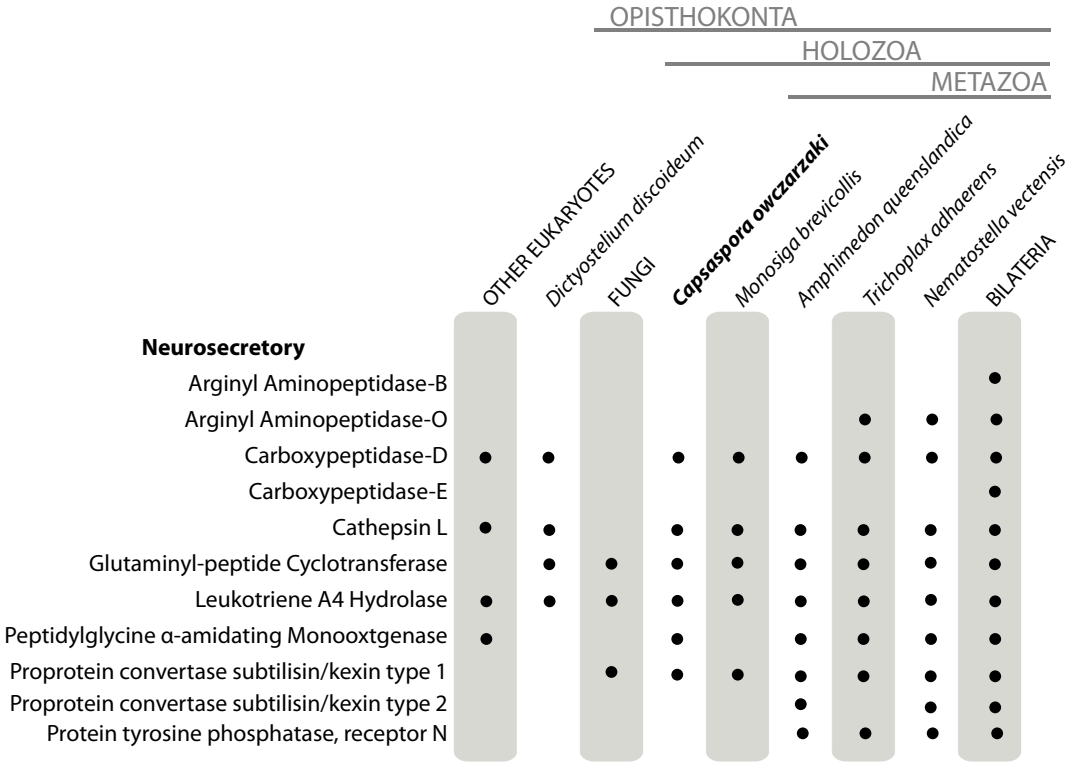
		OPISTHOKONTA							
			HOLOZOA						
				METAZOA					
RNA-binding proteins	OTHER EUKARYOTES	Dictyostelium discoideum	FUNGI	***Cap saspora owczarzaki***	*Monosiga brevicollis*	*Amphimedon queenslandica*	*Trichoplax adhaerens*	*Nematostella vectensis*	BILATERIA
RRM proteins									
elav/Hu		0	0	**0**	1	1	1	2	1-4
musashi		0	0	**0**	0	1	2	3	1-2
dazap1		0	0	**0**	1	2	0	0	1-2
daz/dazl		0	0-1	**1**	0	1	0	1	1
ESRP		0	0	**0**	0	1	1	1	1-2
Spen		0	0	**0**	0	1	1	2	0-1
RBM15		0	0	**0**	0	2	1	1	0-2
bruno		0	0	**1**	1	2	1	3	2-6
others	●	58	53-64	**66**	71	110	78	112	111-207
KH proteins									
FMR1		0	0	**0**	1	1	0	1	1-3
BicC		0	0	**0**	1	1	1	1	1
mex-3		0	0	**0**	0	1	1	1	1-4
others	●	9	10-11	**12**	7	15	12	17	24-34
Dead-box proteins									
vasa		0	0	**0**	0	1	0	3	1
pl10	●	1	1-5	**1**	1	1	1	1	1-2
mov10/armitage a	●	0	0-3	**1**	3	1	0	1	2-3
others	●	48	42-51	**60**	34	69	59	49	59-94
DsRM proteins									
pasha a		0	0	**0**	0	1	0	1	1
drosha a		1	0-1	**0**	1	1	1	2	0-1
dicer a	●	1	1-2	**0**	0	5	5	2	1-2
staufen		0	0	**0**	0	0	0	1	1-2
TARBP2		0	0	**0**	0	1	0	3	1-2
others	●	2	1-3	**2**	3	6	2	4	4-13
Other proteins									
smaug		0	1-2	**1**	0	1	1	1	1-2
muscleblind		0	0	**0**	0	1	1	1	1-4
smn		1	0-1	**1**	1	0	0	0	1
nanos		0	0	**0**	0	1	1	2	1-3
pumilio-A	●	1	2-4	**1**	1	1	1	1	1-2
pumilio-B	●	2	1-6	**1**	0	0	0	0	0
PUF-A	●	1	1	**0**	1	1	0	1	1
PUF-B	●	1	1	**2**	1	0	1	2	1
Exportin-5 a	●	1	1	**1**	1	1	1	1	1
Exportin-1	●	1	1	**1**	1	1	1	1	1
Exportin-T	●	1	1	**1**	1	2	1	1	0-1
piwi a	●	5	0	**0**	0	3	0	3	2-4
argonaute a	●	0	2-7	**0**	0	2	1	2	2-4
Tudor-SN a	●	1	1	**1**	1	1	1	1	1

Supplementary Figure S31. RBP families in unikont lineages

RBP genes are grouped on the basis on their RNA-binding domain (RRM, KH, Dead-box, DsRM, and other RNA-binding domains). Detailed explanation on these families is in[62]. In each group, genes that were not confidently assigned to any of the established family or assigned to other families than those displayed are classified as "others". a, families involved in ncRNA synthesis and functioning.

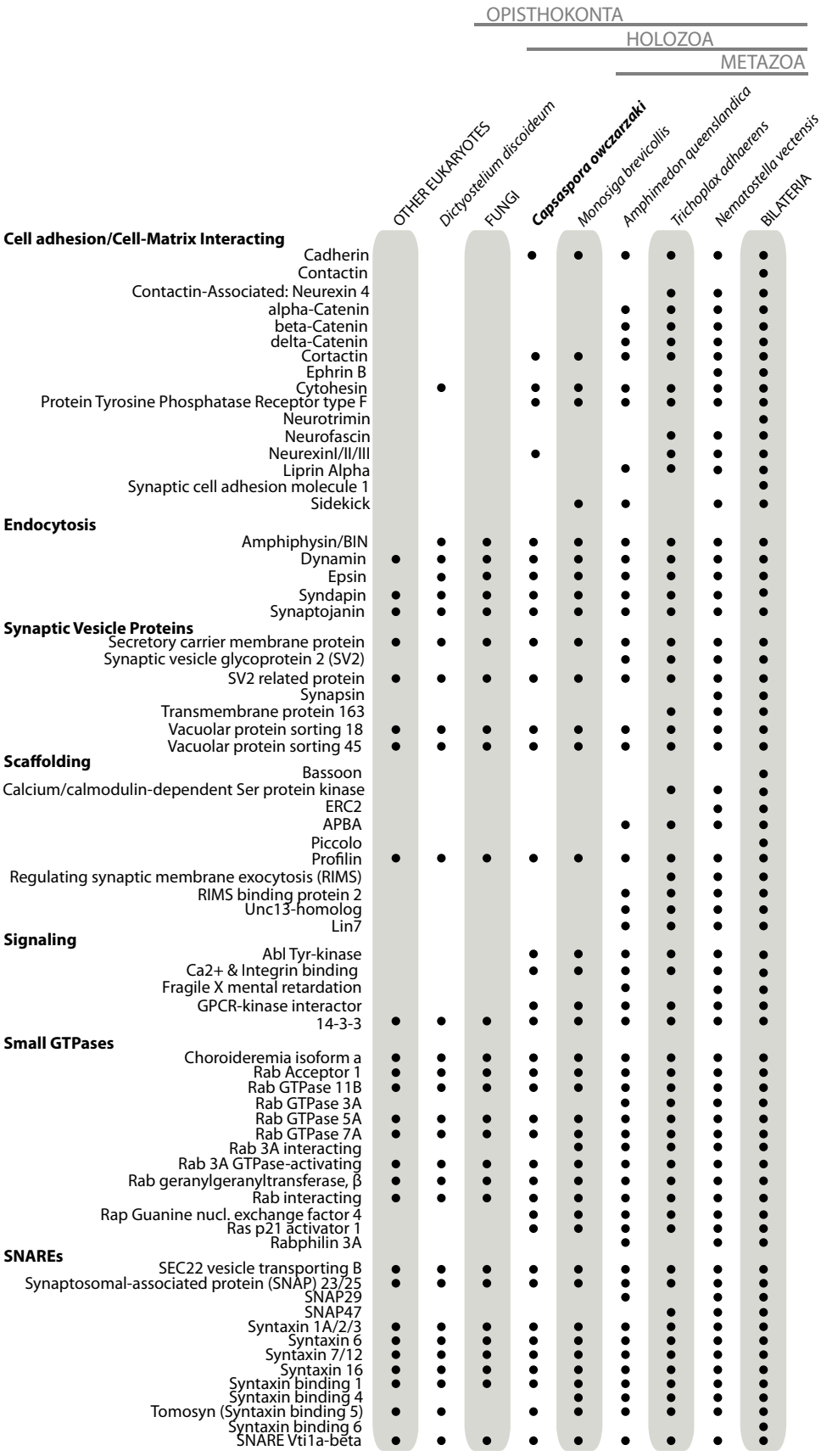


Supplementary Figure S32. Summary of RBP family evolution in eukaryotes
Families that are likely to have been present in eukaryotes before the emergence of unikonts are shown in blue. Those that may have appeared at various stages of unikont evolution are shown in black. Putative family losses are represented by red crosses. For the fungal lineage, only families that are absent in all the examined species are considered as family loss in the Fungi.

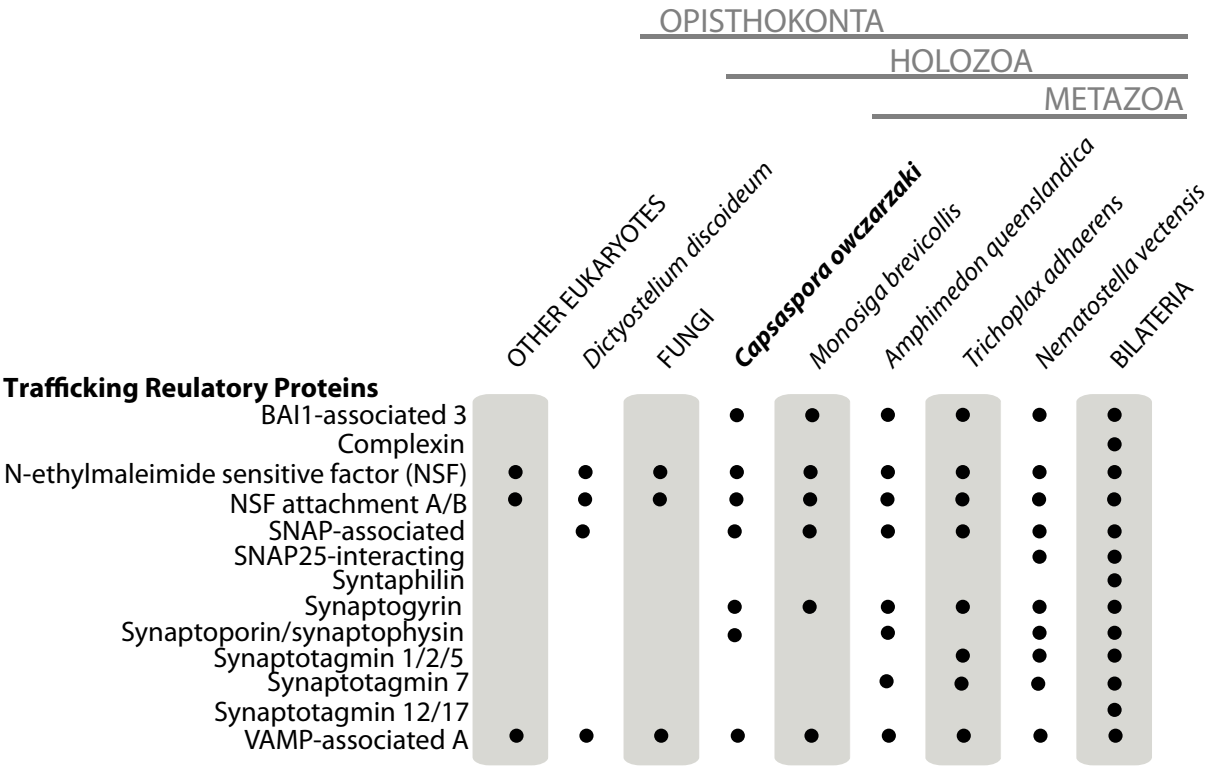


Supplementary Figure S33. Neurosecretion genes

Dots indicate the presence of genes.

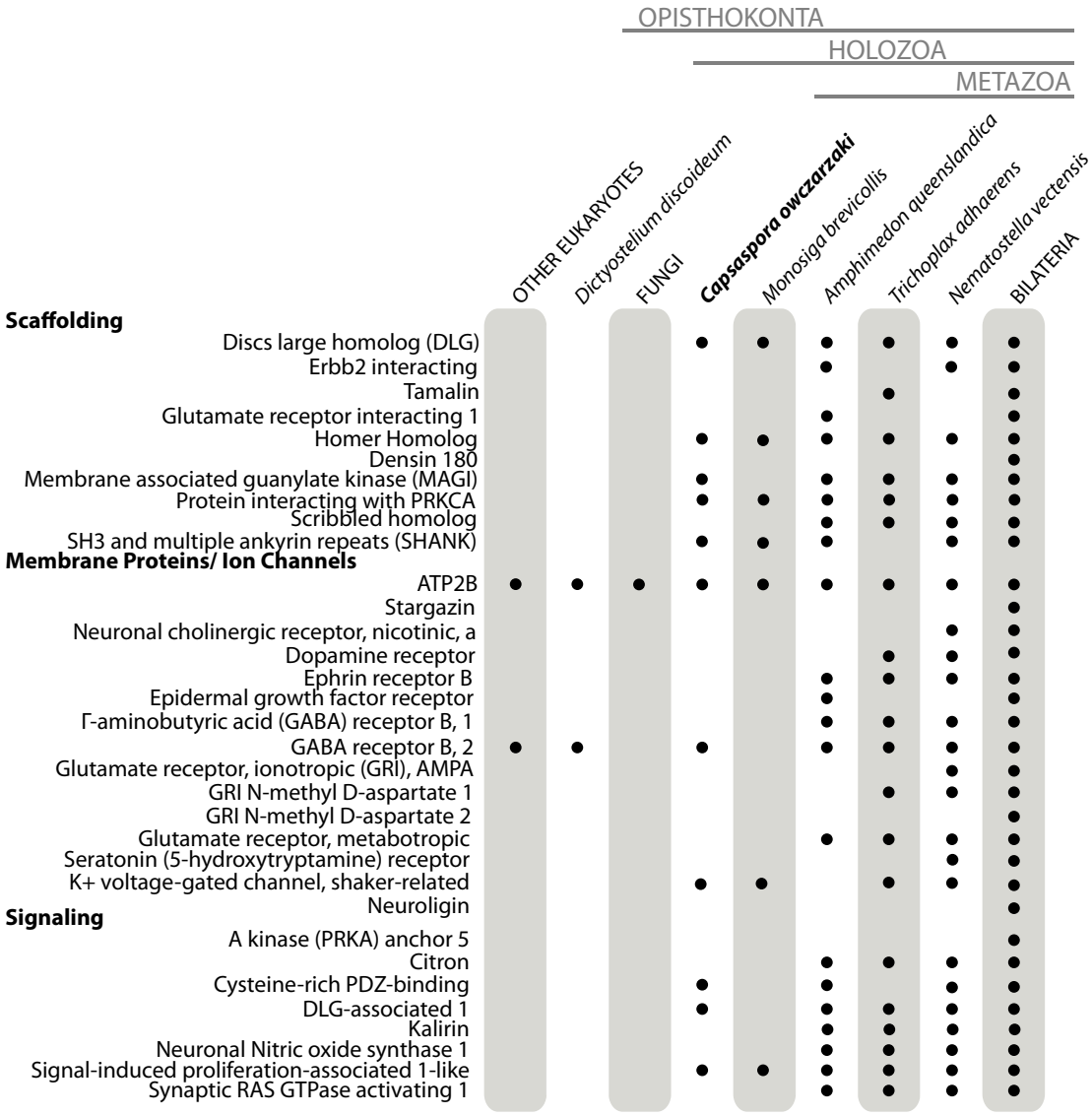


Supplementary Figure S34 continued



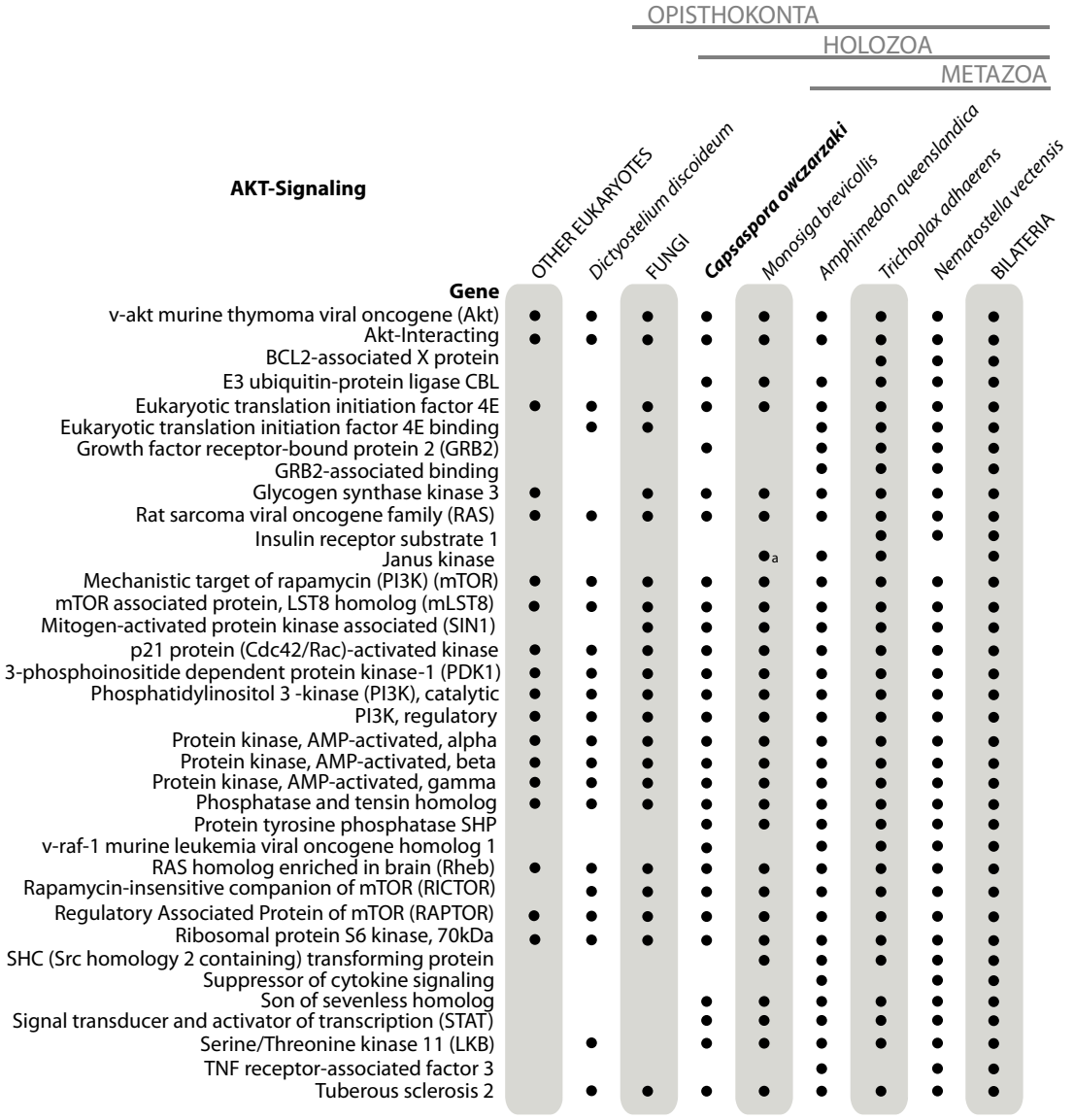
Supplementary Figure S34. Presynaptic genes

Dots indicate the presence of genes.



Supplementary Figure S35. Postsynaptic genes

Dots indicate the presence of genes.



Supplementary Figure S36. Akt signaling genes

Presence and absence of genes encoding Akt signaling components are indicated by dots. a, although a likely homolog of Janus kinase (Jak) is present in *M. brevicollis*, the phylogenetic analysis indicates its close relation to the Spleen tyrosine kinase (Syk) family[20].

Supplementary Tables

Supplementary Table S1. Characterization of transposable element families uncovered in the *C. owczarzaki* genome

LTR Retrotransposons						
Family	Length (bp)	LTR Length (bp)	Target Site Duplication (bp)	Full-Length Element Copy Number	Solo LTR Copy Number	Mean Intra-element LTR Identity (%±SD)
Cocv1	4882	139	5	39	25	99.9±0.23
Cocv2	6208	150	5 and 6	25	10	99.5±0.35
Cocv3	5654	174	5	16	10	100±0.00
Cocv4	6841	80	5	1	16	100
Cocv5	5968	288	5	14	8	99.9±0.001

Non-LTR Retrotransposons			
Family	Length (bp)	Target Site Duplication (bp)	Copy Number
CoL1	6168	13	12
CoL2	6377	12	51
CoL3	6319	n/a	30
CoL4	7141	7	50

Transposons*				
Family	Length (bp)	ITR Length (bp)	Target Site Duplication (bp)	Observed Copy Number (5'/3' ITR)
Cobalt1	2661	59	2	17-34 (17/17)
Cobalt2	2767	83	2	9-18 (9/9)

Cobalt3	2331	27	2	1
CoCACTA1	6987	20	3	18 (9/18)
CoCACTA2	7148	31	3	35 (20/24)
Com1	5477	-	9	19 (15/16)
Com2	5455	30	9	12 (8/5)
Cop1	3139	29	2	28-51 (23/28)
Cop2	3113	29	2	24-45 (20/25)
Cop3	3190	28	2	15-28 (13/15)
Cop4	3195	23	2	19-37 (19/18)
Cop5	3180	27	2	16-31 (16/15)
CoTc1	1652	115	2	41-83 (41/42)
CoTc2	1675	38	2	8-14 (6/8)

* background colored according to the superfamily classification

Supplementary Table S2. General characteristics of mtDNAs from metazoans and their unicellular relatives

Taxon	Size (kbp)	Structure	introns, group	Genetic code
Metazoa				
Homo sapiens	16.6	circular-mapping	-	2
Tethya actinia (demosponge)	19.6	circular-mapping	-	4
Hydra oligactis (cnidarian)	16.3	linear	-	4
Protist relatives of *Metazoa*				
Amoebidium parasiticum	> 200	linear (hundreds)	>23 (group I, II)	4
Capsaspora owczarzaki	196.9	unknown	1 (group I)	4
Monosiga brevicollis	76.6	circular-mapping	4 (group I)	4
Ministeria vibrans	55.0	linear (1 chromosome)	-	4

Supplementary Table S3. List of identified genes in complete mtDNAs of selected metazoans and their unicellular relatives *

Taxon	Complex I-V	Ribosomal proteins	Other proteins	Structural RNAs
H. sapiens	*atp6,8* *cob; cox1,2,3* *nad1-4,4L,5-6*	-	-	*rnl, rns* 22 tRNAs
T. actinia (demosponge)	*atp6,8,9* *cob; cox1,2,3* *nad1-4,4L,5-6*	-	-	*rnl, rns* 25 tRNAs
H. oligactis (cnidarian)	*atp6,8* *cob; cox1,2,3* *nad1-4,4L,5,6*	-	-	*rnl, rns* 2 tRNAs
C. owczarzaki	*atp6,9* *cob; cox1,2,3* *nad1-4,4L,5-6*	*rps3,4,12,13,14,19* *rpl2,5**,6,14,16*	*ccmC,F*	*rnl, rns* 26 tRNAs
M. brevicollis	*atp6,8,9* *cob; cox1,2,3* *nad1-4,4L,5-6*	*rps3,4,8,12,13,14,19* *rpl2,5 **14,16*	*tatC (mttB)*	*rnl, rns* 25 tRNAs
M. vibrans	*atp6,8,9* *cob; cox1,2,3* *nad1-4,4L,5-6*	*rps4,12,13,14,19* *rpl2,6,14,16*	-	*rnl, rns* 25 tRNAs

* Genes and products are: *atp6-9*, ATP-synthase subunits; *cob*, cytochrome b, *cox1-3*, cytochrome oxidase subunits, *nad1-6*, NADH dehydrogenase subunits, *rpl2-16*, large mito-ribosomal subunit proteins; *rps4-19*, small mito-ribosomal subunit proteins; *tatC*, secY-independent transporter; *ccmC, F*, heme delivery and maturation; *rns*, *rnl*, small and large subunit rRNAs.
** weak similarity.

Supplementary Table S4. All Pfam domains gained and lost at selected evolutionary timings (GOs arbitrarily chosen)

Holozoa - gain	
GO:0007155: cell adhesion	Cadherin, Integrin_b_cyt, Integrin_beta
GO:0002376: immune system process	LST1
GO:0007154: cell communication	DSL
GO:0007267: cell-cell signaling	GKAP
GO:0030055: cell-substrate junction	Focal_AT
GO:0005576: extracellular region	A2M_recep, ApoC-I
GO:0019538: protein metabolic process	DUF1908, Hom_end, Peptidase_M2
GO:0016042: lipid catabolic process	HSL_N
GO:0055085: transmembrane transport	ATP-synt_F6
GO:0051082: unfolded protein binding	OmpH
GO:0006810: transport	Band_3_cyto, Pex26, Synaphin, Vitellogenin_N
GO:0044238: primary metabolic process	Fzo_mitofusin, Glyco_transf_6
GO:0008092: cytoskeletal protein binding	ERM
GO:0010468: regulation of gene expression	Churchill, Myc_N, Runt, STAT_alpha, STAT_bind, STAT_int, T-box
GO:0007165: signal transduction	Cbl_N, PI3K_p85B, RBD
GO:0016491: oxidoreductase activity	CI-B14_5a, UCR_6-4kD
GO:0016740: transferase activity	CHGN, PKK
GO:0044425: membrane part	FerB, Macoilin, Sarcoglycan_1, TRAP-delta
GO:0016020: membrane	Ocular_alb
GO:0016043: cellular component organization	P53_tetramer
GO:0016787: hydrolase activity	Rib_hydrolayse, z-alpha
GO:0003677: DNA binding	BAF, P53
GO:0005515: protein binding	PID, PTB
Other categories	PA28_alpha
Unmapped to any GO term	BLVR, BRCA-2_OB3, CENP-M, DDDD, DUF1088, DUF1113, DUF1211, DUF1905, DUF2352, DUF2366, DUF2452, DUF3398, DUF3462, DUF3585, DUF3697, DUF719, DUF737, DUF766, DUF883, DUF902, DUF998, Dsh_C, FSA_C, GIT1_C, HS1_rep, Hemopexin, Hrs_helical, L27, LLGL, L_HGMIC_fpl, Laminin_G_1, Med24_N, Med28, OSTMP1, PEGA, PI3K_1B_p101, PKD, Peptidase_A2B, PriCT_1, Rod_C, SH3BP5, SPOR, ST7, Snurportin1, Strep_SA_rep, Sushi, TPD52, Thiol-ester_cl, Tmemb_18A, YqcI_YcgG, fn2
Holozoa - loss	
GO:0007155: cell adhesion	DUF1881
GO:0006950: response to stress	Barwin, Dehydrin, RTA1, UvdE
GO:0005576: extracellular region	CBM_1, Pectate_lyase
GO:0006259: DNA metabolic process	LAGLIDADG_1
GO:0016070: RNA metabolic process	Intron_maturas2, tRNA_lig_kinase
GO:0006518: peptide metabolic process	Elong-fact-P_C, GCS2, IucA_IucC
GO:0008610: lipid biosynthetic process	CrtC, FAE1_CUT1_RppA
GO:0055085: transmembrane transport	FTR1, K_trans, OPT
GO:0006810: transport	Bac_rhodopsin, Chromate_transp, Form_Nir_trans, Lactate_perm, NQRA
GO:0044238: primary metabolic process	Alginate_lyase, ArabFuran-catal, Bgal_small_N, Cellulose_synt, Glucan_synthase, Glyco_hydro_12, Glyco_hydro_45, Glyco_hydro_6, Glyco_hydro_65N, Glyco_hydro_70, ISN1, LacAB_rpiB, PDEase_II, RhgB_N
GO:0010468: regulation of gene expression	WRKY
GO:0042221: response to chemical stimulus	Erythro_esteras
GO:0016209: antioxidant activity	Peroxidase_2
GO:0008233: peptidase activity	Peptidase_M36
GO:0016740: transferase activity	Chal_sti_synt_C, Chal_sti_synt_N, Chitin_synth_1N, Choline_kin_N, DUF633, mRNA_triPase
GO:0005618: cell wall	Pectinesterase
GO:0044425: membrane part	PsaN, Spore_permease
GO:0016043: cellular component organization	MDM31_MDM32
GO:0004518: nuclease activity	LAGLIDADG_2
GO:0016787: hydrolase activity	Glyco_hydro_53
GO:0016829: lyase activity	PA_decarbox, Terpene_synth_C
Unmapped to any GO term	A_thal_3526, Alginate_lyase2, Amidoligase_2, Asparaginase_II, BLUF, BSP, COBRA, COPI_assoc, CTK3, Cortex-I_coil,

	Cupin_3, Cupin_7, DASH_Dad3, DBD_Tnp_Mut, DIT1_PvcA, DUF1022, DUF1023, DUF1206, DUF1214, DUF1254, DUF1264, DUF1275, DUF1338, DUF1537, DUF1688, DUF1691, DUF1752, DUF1769, DUF1774, DUF1929, DUF1996, DUF2087, DUF2156, DUF2235, DUF2264, DUF2306, DUF2401, DUF2403, DUF2407, DUF2410, DUF2418, DUF2420, DUF247, DUF2470, DUF262, DUF2741, DUF2786, DUF2841, DUF3112, DUF3245, DUF3292, DUF336, DUF3431, DUF3455, DUF3605, DUF3684, DUF3818, DUF521, DUF523, DUF567, DUF939, EutQ, FhuF, GTP_CH_N, Glucoamylase, Glyco_hydro_72, Glyco_transf_36, Glyoxal_oxid_N, HET, Kp4, MKT1_C, MRC1, Metallothio_Euk, Mu-like_Com, NADH-u_ox-rdase, NAS, NPP1, NpwBP, OrfB_Zn_ribbon, PAP1, PG_binding_2, Pec_lyase_C, Peptidase_S58, Peptidase_S64, Pet127, Phi_1, Pho88, Poxvirus_B22R, RHSP, RNA_Me_trans, RPM2, Ran-binding, SKN1, Sec66, Spherulin4, Stm1_N, T5orf172, TIM-br_sig_trns, Tic20, UPF0014, UPF0157, UPF0261, VID27, VTC, X8
Metazoa+Choanoflagellata - gain	
GO:0007155: cell adhesion	Cadherin_pro, Xlink
GO:0002376: immune system process	Somatomedin_B, TNF
GO:0030055: cell-substrate junction	Talin_middle
GO:0005576: extracellular region	Transferrin
GO:0016042: lipid catabolic process	Glyco_hydro_59
GO:0055085: transmembrane transport	ANKH
GO:0006810: transport	DUF1943, Glt_symporter, Selenoprotein_S
GO:0044238: primary metabolic process	Alpha-amylase_N, RbsD_FucU
GO:0008092: cytoskeletal protein binding	Syndecan
GO:0044420: extracellular matrix part	COLFI
GO:0010468: regulation of gene expression	IRF-3, MH2
GO:0007165: signal transduction	GoLoco
GO:0016491: oxidoreductase activity	P4Ha_N
GO:0044425: membrane part	CD225, Dpy19, Tmemb_9
GO:0016020: membrane	MAM
GO:0016787: hydrolase activity	BRK, GGDN
GO:0003676: nucleic acid binding	THAP
GO:0043167: ion binding	PET
Unmapped to any GO term	AcylCoA_DH_N, C1q, CFC, Cadherin_2, Cul7, DAP10, DUF1011, DUF1167, DUF1194, DUF2217, DUF2961, DUF3161, DUF3668, DUF3719, DUF481, DUF729, EB, Excalibur, Fer4_3, I-set, L27_1, Laminin_N, MAR_sialic_bdg, MRP-S22, Mab-21, Mucin, PMG, Plexin_cytopl, RBD-FIP, Reeler, VASP, VPEP, YcgR_2, ig
Metazoa+Choanoflagellata - loss	
GO:0004888: transmembrane signaling receptor activity	PAS_2, PHY
GO:0019538: protein metabolic process	Hom_end_hint
GO:0006810: transport	Brr6_like_C_C, FUSC, GtrA, PDR_CDR
GO:0010468: regulation of gene expression	Zn_clus
GO:0042221: response to chemical stimulus	ALMT
GO:0016491: oxidoreductase activity	DUF1729, GlutR_N
GO:0016740: transferase activity	Transferase
GO:0016787: hydrolase activity	PriCT_2
GO:0016874: ligase activity	GSIII_N
Unmapped to any GO term	Bot1p, CRT-like, CVNH, CcmH, CotH, D5_N, DUF1212, DUF1237, DUF1304, DUF1764, DUF2015, DUF2201, DUF2283, DUF231, DUF2343, DUF2421, DUF2804, DUF2823, DUF2834, DUF3712, DUF3722, DUF3815, DUF45, DUF457, DinB, Gti1_Pac2, HPP, PDZ_1, QCR10, RNA_lig_T4_1, SUA5, Spo7, Suc_Fer-like, TRP, Tir_receptor_C, Velvet, XYPPX, Ytp1
Metazoa - gain	
GO:0007155: cell adhesion	C2-set, NIDO, TSP_C
GO:0010941: regulation of cell death	Bcl-2, DED, FAIM1
GO:0050793: regulation of developmental process	Noggin

GO:0007267: cell-cell signaling	HH_signal
GO:0005102: receptor binding	Somatostatin, TGF_beta, wnt
GO:0022414: reproductive process	Sp38
GO:0006950: response to stress	RuvB_C
GO:0016567: protein ubiquitination	USP8_interact
GO:0044217: other organism part	Herpes_gI
GO:0005576: extracellular region	IGFBP, Uteroglobin
GO:0004888: transmembrane signaling receptor activity	Ephrin_lbd, HRM
GO:0006259: DNA metabolic process	DNA_pol3_alpha, DNA_pol3_chi
GO:0016070: RNA metabolic process	SirA
GO:0019538: protein metabolic process	ADAM_CR, DMPK_coil, MRP-S23, PDCD9, Transglut_N, zf-C4_ClpX
GO:0008610: lipid biosynthetic process	FA_synthesis, PhaC_N
GO:0032774: RNA biosynthetic process	Rho_RNA_bind
GO:0055085: transmembrane transport	OAD_gamma
GO:0006810: transport	DuoxA, Hemocyanin_M, Invas_SpaK, MotA_ExbB, Na_K-ATPase, SBP_bac_7, Secretin, TonB, oligo_HPY
GO:0044238: primary metabolic process	AceK
GO:0031012: extracellular matrix	ADAM_spacer1, Lamprin
GO:0008092: cytoskeletal protein binding	Tropomodulin
GO:0010468: regulation of gene expression	Ets, HNF-1_N, HTH_DeoR, Hairy_orange, MH1, NCD1, PCAF_N, Pou, RHD, SCAN, TAFH, zf-C2HC, zf-C4
GO:0007165: signal transduction	Death, MNNL, RanGAP1_C, SOCS_box
GO:0019207: kinase regulator activity	CDK5_activator
GO:0000988: protein binding transcription factor activity	CBF_beta
GO:0016491: oxidoreductase activity	COX6C, Lysyl_oxidase, MmoB_DmpM, NADH_oxidored, PdxA
GO:0016740: transferase activity	HSNSD, Preseq_ALAS, WIF
GO:0044427: chromosomal part	SCP-1
GO:0044425: membrane part	Anth_Ig, FTSW_RODA_SPOVE, FerA
GO:0016020: membrane	Lamp
GO:0005634: nucleus	HARP, Ski_Sno, c-SKI_SMAD_bind
GO:0031967: organelle envelope	LEM
GO:0019222: regulation of metabolic process	GFRP, Geminin, PP1_inhibitor
GO:0004518: nuclease activity	Herpes_alk_exo
GO:0016787: hydrolase activity	PP2C_C, RIG-I_C-RD
GO:0016874: ligase activity	DUF3590
GO:0008152: metabolic process	FlpD, PdxJ
GO:0051540: metal cluster binding	FeS
GO:0003677: DNA binding	BESS, IRF
GO:0005515: protein binding	Ldl_recept_a, Sema, VWC
GO:0043167: ion binding	zf-dskA_traR
Other categories	Cuticle_1, NPV_P10
Unmapped to any GO term	7TM_GPCR_Srbc, 7TM_GPCR_Srx, AF-4, AMOP, APP_amyloid, Autotransporter, Avidin, BCL_N, BEN, BNIP2, Beta-TrCP_D, BiPBP_C, BrkDBD, C1-set, C2-set_2, CABIT, CALCOCO1, CDK2AP, CTNNB1_binding, CXCXC, Ca_chan_IQ, Calponin, Cas_Cas1, Caveolin, CbbQ_C, Cor1, CtIP_N, CusF_Ec, Cuticle_3, DAG1, DAZAP2, DBD_Tnp_Hermes, DUF1016, DUF1041, DUF1208, DUF1280, DUF1387, DUF1456, DUF1520, DUF1735, DUF1758, DUF2051, DUF2216, DUF2353, DUF2668, DUF3105, DUF3244, DUF3394, DUF3447, DUF3469, DUF3497, DUF3504, DUF3512, DUF3513, DUF3518, DUF3524, DUF3534, DUF3643, DUF3695, DUF898, DUF948, DZF, Daxx, DctQ, DsrC, E3_UbLigase_EDD, EGF_MSP1_1, EIF4E-T, Endonuclease_7, Exonuc_V_gamma, FA, FAST_2, FCD, FEZ, GTF2I, GerPC, HEPN, Hemocyanin_C, ICAP-1_inte_bdg, IF2_assoc, IRF-2BP1_2, Integrin_alpha2, JHBP, Jnk-SapK_ap_N, KSHV_K1, Lipoprotein_18, MOZART2, MacB_PCD, Med25, Mitoc_L55, NOPS, Nebulin, Neuralized, NfI_DNAbd_pre-N, Nrf1_DNA-bind, OCIA, OstA, PIP49_C, PP1c_bdg, PaaA_PaaC, Peptidase_A17, Phage_30_3, Phage_head_chap, Phospho_p8, RBB1NT, RDD, RGM_C, Receptor_2B4, SERTA, STOP, SUFU, SUFU_C, SapA, Sec31, Serine_rich, Smoothelin, TET_Cys_rich, TET_DSBH, TF_AP-2, TIL, Terminase_GpA, Thyroglobulin_1, Tissue_fac,

	Tmemb_55A, Tmemb_cc2, UPF0561, V-set, V-set_CD47, VWA_N, Xylo_C, YajC, ZU5, ZapA, gpUL132, plasmid_Toxin, zf-nanos
Metazoa - loss	
GO:0006950: response to stress	OsmC
GO:0005694: chromosome	Topo-VIb_trans
GO:0006259: DNA metabolic process	S1-P1_nuclease
GO:0016070: RNA metabolic process	PARP_regulatory, tRNA_lig_CPD
GO:0019538: protein metabolic process	Leu_Phe_trans
GO:0006810: transport	RhaT
GO:0044238: primary metabolic process	Alpha-L-AF_C, Chalcone, Glyco_hydro_43, HisG_C, PDT, PEPCK_ATP, PRA-CH, Pantoate_ligase
GO:0008233: peptidase activity	DUF3586
GO:0016491: oxidoreductase activity	Desulfoferrodox, Shikimate_dh_N
GO:0016740: transferase activity	ATP_transf, Init_tRNA_PT
GO:0044425: membrane part	Cons_hypoth698
GO:0016829: lyase activity	ADC, Anth_synt_I_N, DHquinase_I
GO:0008152: metabolic process	CM_2
Unmapped to any GO term	Act-Frag_cataly, DUF1365, DUF1513, DUF1765, DUF1993, DUF2009, DUF3228, DUF3295, DUF3336, DUF707, DUF711, DUF72, DUF952, DUF962, Glyco_hydro_101, Glyco_hydro_32N, HR_lesion, LtrA, Lyase_8_N, Malectin_like, Mating_C, NCA2, PH_2, ParBc_2, PqiA, Saccharop_dh_N, ThuA, UPF0052, UPF0126, VIT1, WLM

Supplementary Table S5. Significantly gained GO terms and included Pfam domains

Clade	GO Term (Topology-Weighted)	p-value	Pfam domains included
Holozoa	Signal transducer activity	6.7e-7	Integrin_b_cyt, RBD, Integrin_beta, STAT_int, Cbl_N, STAT_bind, STAT_alpha, Focal_AT
	Transcription regulatory region DNA binding*	1.3e-5	Runt, P53, T-box, Myc_N, STAT_int, STAT_bind, STAT_alpha
	Integrin mediated signaling pathway	3.7e-4	Integrin_b_cyt, Integrin_beta
Fungi	*No significant GO terms*		-
Metazoa + *M. brevicollis*	Immune response	6.5e-4	TNF, Somatomedin_B
Metazoa	Extracellular region	3.3e-5	IGFBP, wnt, Herpes_gI, Sp38, Lamprin, TSP_C, Uteroglobin, ADAM_spacer1, Somatostatin
	Regulation of apoptosis	5.0e-4	Bcl-2, DED, FAIM1
	Regulation of transcription, DNA-dependent**	9.4e-4	zf-C4, PCAF_N, MH1, HTH_DeoR, TAFH, RHD, HNF-1_N, Hairy_orange, SCAN, Ets, zf-C2HC, Pou, NCD1
M. brevicollis	Hydrolyase activity	5.6e-4	GD_AH_C, UxuA
C. owczarzaki	*No significant GO terms*		-

* a similar GO term (regulation of transcription, DNA-dependent) contains the Pfam domain Churchill, but not P53, and is not shown here because of the relatively high p-value (0.007).

** a similar GO term (transcription regulatory region DNA binding) is also highly significant (p = 4.7e-4) but does not include the Pfam domains PCAF_N, MH1, HNF-1_N, Hairy_orange, and NCD1.

Supplementary Table S6. Significantly lost GO terms and included Pfam domains

Clade	GO Term (Topology-Weighted)	p-value	Pfam domains included
Holozoa	Carbon-oxygen lyase activity, acting on polysaccharides	3.9e-5	Pectate_lyase, Alginate_lyase, RhgB_N
	Glucan biosynthetic process	9.5e-5	Cellulose_synt, Glucan_synthase, Glyco_hydro_70
	Peptide biosynthetic process	5.1e-4	IucA_IucC, Elong-fact-P_C, GCS2
	Hydrolase activity (O-glycosyl bonds)	5.5e-4	Glyco_hydro_45, CBM_1, ArabFuran-catal, Glyco_hydro_53, Glyco_hydro_12, Glyco_hydro_6, Bgal_small_N
Fungi	Multi-organism process	7.9e-5	Endotoxin_N, Toxin_2, Anemone_cytotox, Fusion_gly, T4SS-DNA_transf, Cytadhesin_P30, Adeno_shaft, ADPrib_exo_Tox, Toxin_R_bind_N, Triabin
	Negative regulation of hydrolase activity	1.3e-4	Cystatin, A2M_N, WAP, Antistasin, Kunitz_BPTI, Kunitz_legume, potato_inhibit
	Thylakoid	1.8e-4	CytB6-F_Fe-S, PsaA_PsaB, PsbI, PsbJ, PSI_PsaJ, Apocytochr_F_C
	Transcription regulatory region DNA ninding	2.4e-4	CarD_TRCF, TSC22, HTH_5, HTH_8, HTH_1, Hormone_recep, AsnC_trans_reg, K-box, FUR, Sigma70_r1_2, LacI, GerE, Sigma70_r2, Sigma70_r4, GntR, MarR
	Defense response	6.0e-4	Gamma-thionin, Antimicrobial19, Thionin, KMP11, TIR, Triabin
Metazoa + *M. brevicollis*	G-protein coupled photoreceptor activity	2.9e-5	PHY, PAS_2
Metazoa	*No significant GO terms*		-
M. brevicollis	Extracellular region	4.4e-9	Motilin_assoc, A2M_recep, Toxin_1, Toxin_2, DUF290, WAP, Antimicrobial_2, ApoL, CBM_5_12, Peptidase_M10, Aerolysin, Transglycosylas, PLA2G12, Dickkopf_N, Glypican, Hydrophobin_2, A2M_comp, Galanin, Stanniocalcin, A_deaminase_N, Glyco_hydro_46, KRE9, ApoC-I, Crust_neurohorm, Glyco_hydro_67M, CBM_14, Toxin_R_bind_N, Laminin_B, PepSY
	Regulation of transcription, DNA-dependent	2.3e-6	HIRA_B, Cenp-B_dimeris, NOT2_3_5, QLQ, Hira, BRF1, YL1, Med13_C, CarD_TRCF, TSC22, Med15_fungi, CHDCT2, NusB, K-box, FUR, Sigma70_r1_2, HALZ, GerE, CBFB_NFYA, Pencillinase_R, Sigma70_r2, Sigma70_r4, Hap4_Hap_bind, Crp, CT20, FLO_LFY, Bvg_acc_factor, Myc_N, Med8, Med9, Med1, Med31, STAT_alpha, Tfb4, Trans_reg_C, Sigma70_r4_2, PAX, STAT_int, Med15, Med17, Med18, Med19, Med12, Med11, HTH_5, HTH_6, HTH_8, Med20, HTH_1, NCD2, Hormone_recep, AsnC_trans_reg, T-box, Churchill, DMAP1, Beta-trefoil,

			GntR, Nuc_rec_co-act, CSD, Runt, TFIIF_alpha, MarR
	Nucleoplasm	1.3e-4	TFIID_20kDa, Med13_C, Med15_fungi, CT20, Med8, Med9, Med1, Med31, TAFII55_N, TFIIA_gamma_N, TFIIA_gamma_C, TFIIE_beta, TAF4, Med15, Med17, Med18, Med19, Med12, Med11, Med29, Med20, ELL
	Developmental process	1.8e-4	DIX, Dishevelled, LIM_bind, Sina, Spore_GerAC, Dickkopf_N, Aegerolysin, MinC_C, HCR, Churchill, LST1
	Cellular cell wall organization or biogenesis	4.1e-4	Transgly, XG_Ftase, Dala_Dala_lig_N, Glyco_hydro_26, Extensin_2, Transpeptidase, KRE9
	Racemase and epimerase activity	8.5e-4	Epimerase_2, NanE, DUF718, Pro_racemase, Ribul_P_3_epim, Polysacc_syn_2C, Ala_racemase_C, C5-epim_C
C. owczarzaki	Extracellular region	1.0e-6	Dickkopf_N, Motilin_assoc, Toxin_1, Toxin_2, DUF290, WAP, Glypican, Hydrophobin_2, Galanin, Stanniocalcin, Glyco_hydro_46, KRE9, ADPrib_exo_Tox, Antimicrobial_2, ApoL, CBM_5_12, Peptidase_M10, Aerolysin, Transglycosylas, Crust_neurohorm, Glyco_hydro_67M, CBM_14, Lyase_8, Toxin_R_bind_N, Laminin_B, PepSY
	Receptor binding	1.4e-6	Motilin_assoc, Galanin, NGF, Stanniocalcin, Rapsyn_N, Fibrinogen_C, Crust_neurohorm, TGFb_propeptide, Toxin_R_bind_N, Nuc_rec_co-act, PTN_MK_C, IRS
	Reproduction	1.4e-6	Vac_Fusion, Terminase_5, VOMI, Aegerolysin, MEA1, Fusion_gly, DUF1898, Flavi_DEAD, Phage_T7_Capsid, Herpes_Helicase, Adeno_IVa2, DUF904
	Multicellular organismal process	8.8e-6	Dickkopf_N, 7tm_6, VOMI, Aegerolysin, MEA1, DUF1898, Rapsyn_N, Adeno_shaft, DIX, Dishevelled, Toxin_R_bind_N, Sina, Synapsin
	Cell wall organization biogenesis	2.3e-5	Transgly, Glyco_hydro_81, XG_FTase, KRE9, Dala_Dala_lig_N, Glyco_hydro_26, Extensin_2, Glyco_hydro_19, Transpeptidase, Phage_lysozyme, Glyco_hydro_67M
	Pathogenesis	2.7e-5	Endotoxin_N, Toxin_2, Cytadhesin_P30, Adeno_shaft, ADPrib_exo_Tox, Aerolysin, Toxin_R_bind_N, CDtoxinA
	Viral reproduction	5.5e-5	Vac_Fusion, Terminase_5, Pox_A_type_inc, Fusion_gly, Flavi_DEAD, Herpes_gp2, Phage_T7_Capsid, Herpes_Helicase, Adeno_IVa2
	Nickel ion binding	3.8e-4	UreF, UreD, HypA, Urease_beta,

			Urease_gamma, Urease_alpha
	Thylakoid	3.8e-4	CytB6-F_Fe-S, PsaA_PsaB, PsbI, PsbJ, PSI_PsaJ, Apocytochr_F_C
	External encapsulating structure	6.2e-4	FlaA, Transgly, Hydrophobin_2, DUF1034, Dala_Dala_lig_N, TctC, SBP_bac_3

Supplementary Table S7. Metazoan-specific Pfam domains inferred by the canonical parsimony within the Opisthokonta (GOs arbitrarily chosen)*

GO:0007155: cell adhesion	Adeno_shaft, C2-set, Cytadhesin_P30, Laminin_B, NIDO, TSP_3, TSP_C
GO:0010941: regulation of cell death	Bcl-2, CARD, DED, FAIM1, TRADD_N
GO:0050793: regulation of developmental process	Noggin
GO:0008219: cell death	FAST_1
GO:0007267: cell-cell signaling	HH_signal
GO:0005102: receptor binding	Crust_neurohorm, NGF, Nuc_rec_co-act, PTN_MK_C, Somatostatin, TGF_beta, TGFb_propeptide, Toxin_R_bind_N, wnt
GO:0022414: reproductive process	Fusion_gly, Sp38, VOMI
GO:0009405: pathogenesis	Endotoxin_N, Toxin_2
GO:0006950: response to stress	Adenine_glyco, Antimicrobial19, Gamma-thionin, KMP11, Methyltransf_1N, Pur_DNA_glyco, RuvB_C, RuvC, TRCF, Thionin, Triabin, UPF0081, UvrC_HhH_N, potato_inhibit
GO:0016567: protein ubiquitination	USP8_interact
GO:0044217: other organism part	Herpes_gI
GO:0005694: chromosome	DNA_gyraseA_C, DNA_gyraseB_C
GO:0005576: extracellular region	Antimicrobial_2, ApoL, Dickkopf_N, IGFBP, Toxin_1, Transglycosylas, Uteroglobin, WAP
GO:0004888: transmembrane signaling receptor activity	7tm_6, Ephrin_lbd, HRM, Lig_chan, Lig_chan-Glu_bd, SRCR, TarH
GO:0006259: DNA metabolic process	DNA_pol3_alpha, DNA_pol3_beta, DNA_pol3_chi, DNA_pol_B_2, DnaB, Exonuc_VII_S, Integrase_DNA, RAG2, Rep_3, zf-CHC2
GO:0016070: RNA metabolic process	SPOUT_MTase, SirA, tRNA-synt_2e
GO:0019538: protein metabolic process	ADAM_CR, DMPK_coil, FBA, Glyco_transf_29, MRP-S23, PDCD9, Peptidase_C39, Peptidase_M32, Peptidase_S11, Peptidase_U32, Ribosomal_L25p, Ribosomal_S20p, Ribosomal_S3_N, Sina, Transglut_N, Trigger_C, Trigger_N, VKG_Carbox, zf-C4_ClpX
GO:0008144: drug binding	PBP_dimer, Transpeptidase
GO:0006518: peptide metabolic process	GSH-S_ATP
GO:0016042: lipid catabolic process	Phospholip_A2_1
GO:0008610: lipid biosynthetic process	CTP_transf_3, CmcI, DXP_redisom_C, DXP_reductoisom, FA_synthesis, GcpE, LAB_N, LYTB, LpxK, PhaC_N, Polysacc_syn_2C
GO:0032774: RNA biosynthetic process	Hox9_act, RNA_pol_A_CTD, Rho_RNA_bind, Sigma70_r1_2, Sigma70_r2, Sigma70_r4
GO:0055085: transmembrane transport	ATP-synt_B, ATP-synt_DE, GntP_permease, OAD_gamma, Peripla_BP_2, YMF19
GO:0006810: transport	ACR_tran, ASC, Anemone_cytotox, BCCT, BPD_transp_1, BPD_transp_2, Bac_globin, BicD, DuoxA, H_PPase, Hemocyanin_M, Invas_SpaK, LysE, MotA_ExbB, MttA_Hcf106, NSF, Na_K-ATPase, Na_Pi_cotrans, SBP_bac_1, SBP_bac_3, SBP_bac_5, SBP_bac_7, SecA_SW, Secretin, Synapsin, TOBE_2, TonB, oligo_HPY
GO:0044238: primary metabolic process	AceK, Alpha-amyl_C2, CBM_11, DDE_Tnp_ISL3, DapB_C, FA_desaturase_2, GPI, Glyco_hydro_19, Glyco_hydro_77, Glycos_transf_N, Lip_A_acyltrans, MethyltransfD12, NanE, PFL, PUD, PYNP_C, Polysacc_synt, Sucrose_synth, Transgly, Varsurf_PPLC
GO:0031012: extracellular matrix	ADAM_spacer1, Glypican, Lamprin
GO:0007017: microtubule-based process	FOP_dimer, Tektin
GO:0008092: cytoskeletal protein binding	Tropomodulin
GO:0010468: regulation of gene expression	AsnC_trans_reg, Bvg_acc_factor, CHDCT2, CarD_TRCF, Cenp-B_dimeris, Ets, FLO_LFY, FUR, GerE, GntR, HALZ, HNF-1_N, HTH_1, HTH_5, HTH_8, HTH_DeoR, Hairy_orange, Hormone_recep, K-box, MH1, MarR, NCD1, NusB, PCAF_N, Pou, RHD, SCAN, TAFH, TSC22, zf-C2HC, zf-C4
GO:0007165: signal transduction	APC_crr, Death, HisKA_3, His_kinase, MCPsignal, MNNL, RanGAP1_C, SOCS_box
GO:0042221: response to chemical stimulus	MecA_N
GO:0009416: response to light stimulus	NPH3
GO:0019207: kinase regulator activity	CDI, CDK5_activator

GO:0008233: peptidase activity	Hepsin-SRCR, Peptidase_M66, Peptidase_S49_N
GO:0000988: protein binding transcription factor activity	CBF_beta
GO:0030234: enzyme regulator activity	Kunitz_legume, P-II, PMEI
GO:0016491: oxidoreductase activity	C1_3, COX6C, COX7a, CytB6-F_Fe-S, DsbB, EKR, IDH, Lysyl_oxidase, MmoB_DmpM, NADH_oxidored, PPO1_DWL, PPO1_KFDV, PQQ_C, PdxA, Ring_hydroxyl_A, T4_deiodinase, VDE
GO:0016740: transferase activity	CAT, DNA_pol3_gamma3, DUF299, FTCD, HSNSD, Mec-17, Methyltrans_SAM, PTA_PTB, Preseq_ALAS, WIF, YkuD
GO:0044427: chromosomal part	SCP-1
GO:0044425: membrane part	Anth_Ig, Apocytochr_F_C, CD20, EzrA, FTSW_RODA_SPOVE, FerA, Herpes_gp2, Multi_Drug_Res, PSI_PsaJ, Pox_P21, PsaA_PsaB, PsbI, PsbJ, SDH_sah
GO:0016020: membrane	Bac_Ubq_Cox, Cache_1, Cys_rich_FGFR, Cytochrom_C_asm, FecCD, FtsX, Herpes_gE, Lamp, Spore_GerAC, Surf_Ag_VNR, T4SS-DNA_transf
GO:0005634: nucleus	AKAP95, DUF1143, EIN3, Gemin6, HARP, KNOX1, NOA36, SAND, Ski_Sno, c-SKI_SMAD_bind
GO:0031967: organelle envelope	LEM
GO:0042597: periplasmic space	FlaA, LTXXQ, TctC
GO:0044422: organelle part	Med29
GO:0005615: extracellular space	DUF290
GO:0016043: cellular component organization	DUF904, Extensin_2, MinC_C
GO:0019222: regulation of metabolic process	GFRP, Geminin, PP1_inhibitor
GO:0007275: multicellular organismal development	DIX, Dishevelled
GO:0048870: cell motility	Flg_bb_rod
GO:0004518: nuclease activity	5_3_exonuc, 5_3_exonuc_N, Endonuclease_1, Herpes_alk_exo
GO:0016787: hydrolase activity	ATP_Ca_trans_C, Acid_phosphat_B, ChitinaseA_N, DUF3357, Destabilase, PDEase_I_N, PP2C_C, PTE, Peptidase_S7, RIG-I_C-RD, XendoU, YcfA
GO:0016829: lyase activity	Ectoine_synth, Pec_lyase_N
GO:0016874: ligase activity	DUF3590, GAD
GO:0008152: metabolic process	CmcH_NodU, CobN-Mg_chel, FTCD_C, FlpD, PdxJ, ThiC, UPF0051
GO:0051540: metal cluster binding	FeS
GO:0009055: electron carrier activity	Rubredoxin
GO:0003677: DNA binding	BESS, Bac_DNA_binding, IRF, TetR_N, zf-CXXC
GO:0003676: nucleic acid binding	Agenet, CRS1_YhbY, IN_DBD_C, SmpB
GO:0005515: protein binding	CCT_2, Fz, Ldl_recept_a, Sema, VWC
GO:0043167: ion binding	HemolysinCabind, PhosphMutase, zf-CW, zf-dskA_traR
GO:0019028: viral capsid	Baculo_PEP_C, Gag_p10, NPV_P10
Other categories:	AXH, Abi_HHR, Cuticle_1, DUF1967, Lipocalin, Lipocalin_2, SBP56
Unmapped to any GO term:	2_5_RNA_ligase, 3-PAP, 4HB_MCP_2, 53-BP1_Tudor, 7TMR-DISM_7TM, 7TM_GPCR_Srbc, 7TM_GPCR_Srsx, 7TM_GPCR_Srx, A2L_zn_ribbon, A2M_N_2, AF-4, AMOP, APP_amyloid, ARL2_Bind_BART, Angiomotin_C, ApbE, ApoB100_C, AraC_E_bind, AraC_N, Aurora-A_bind, Autotransporter, Avidin, Avirulence, Axin_b-cat_bind, BAT2_N, BCA_ABC_TP_C, BCL_N, BEN, BNIP2, Baculo_LEF5_C, Barstar, BaxI_1, Beta-TrCP_D, BiPBP_C, Big_1, BrkDBD, C1-set, C1_2, C2-set_2, CABIT, CALCOCO1, CBM_X, CDK2AP, CHRD, COXG, CRAM_rpt, CTNNB1_binding, CXCXC, CaATP_NAI, Ca_chan_IQ, Calmodulin_bind, Calponin, Caps_synth, Cas_Cas1, Caveolin, CbbQ_C, Chlam_PMP, Cna_B, Collar, ComX, Cor1, CorC_HlyC, CpeT, CreA, Crisp, Crystall, CtIP_N, CusF_Ec, Cuticle_3, CxxC_CxxC_SSSS, DAG1, DAZAP2, DBD_Tnp_Hermes, DELLA, DM13, DRTGG, DUF1016, DUF1041, DUF1074, DUF1076, DUF1086, DUF1087, DUF1096, DUF111, DUF1118, DUF1120, DUF1152, DUF1193, DUF1208, DUF1223, DUF1255, DUF1280, DUF1289, DUF1292, DUF1294, DUF13, DUF1313, DUF1356, DUF137, DUF1387, DUF1390,

DUF1409, DUF1416, DUF1421, DUF1450, DUF1456, DUF149, DUF1493, DUF150, DUF1517, DUF1520, DUF1524, DUF1557, DUF1601, DUF162, DUF1639, DUF1664, DUF1668, DUF1685, DUF1704, DUF1731, DUF1732, DUF1735, DUF1737, DUF1758, DUF177, DUF179, DUF1794, DUF1816, DUF1863, DUF1873, DUF1918, DUF1927, DUF1935, DUF1949, DUF1986, DUF1997, DUF2051, DUF2053, DUF2064, DUF208, DUF2118, DUF2135, DUF2181, DUF2185, DUF220, DUF2207, DUF2216, DUF2252, DUF2256, DUF2322, DUF2345, DUF2353, DUF2359, DUF2368, DUF2379, DUF2448, DUF2475, DUF2487, DUF249, DUF26, DUF2604, DUF2647, DUF2668, DUF2691, DUF2738, DUF2750, DUF2807, DUF2817, DUF288, DUF2920, DUF2993, DUF2997, DUF3007, DUF3072, DUF3079, DUF3105, DUF3133, DUF3139, DUF3148, DUF316, DUF3170, DUF3244, DUF3248, DUF3250, DUF3252, DUF3276, DUF3335, DUF3394, DUF3411, DUF3420, DUF3447, DUF3458, DUF3469, DUF348, DUF3497, DUF3504, DUF3512, DUF3513, DUF3518, DUF3524, DUF3534, DUF3583, DUF3592, DUF3598, DUF3641, DUF3643, DUF3648, DUF3656, DUF3677, DUF3695, DUF37, DUF3848, DUF393, DUF43, DUF445, DUF477, DUF490, DUF501, DUF506, DUF553, DUF561, DUF566, DUF569, DUF577, DUF581, DUF599, DUF606, DUF615, DUF622, DUF626, DUF629, DUF630, DUF639, DUF640, DUF641, DUF662, DUF668, DUF677, DUF702, DUF705, DUF716, DUF724, DUF755, DUF781, DUF796, DUF800, DUF819, DUF828, DUF839, DUF849, DUF853, DUF885, DUF892, DUF898, DUF948, DUF955, DUF972, DUF98, DZF, DZR, Daxx, DbpA, DctM, DctQ, Di19, Dicty_CTDC, Dicty_spore_N, DnaB_2, DnaI_N, DrsE, DsrC, Dynein_IC2, E3_UbLigase_EDD, EAL, EFP_N, EGF_MSP1_1, EIF4E-T, ENT, E_Pc_C, Endonuc-dimeris, Endonuclease_7, Exonuc_V_gamma, FA, FAST_2, FBD, FCD, FEZ, FIST_C, FLYWCH, FNIP, FTH, Fer4_5, Fibrinogen_aC, FlaC_arch, Flavodoxin_NdrI, Flg_hook, FliB, FmrO, FrhB_FdhB_N, FtsZ_C, Ftsk_gamma, GABP-alpha, GAGA, GAGA_bind, GNAT_acetyltran, GRP, GSPII_F, GTF2I, GerPC, GlcNAc, Glutaredoxin2_C, GvpG, HDOD, HEPN, HTH_18, HTH_20, HTH_OrfB_IS605, HTH_Tnp_Tc3_1, Hemocyanin_C, Herpes_UL45, HipA_C, HpaB_N, HxlR, ICAP-1_inte_bdg, IF2_assoc, IFP_35_N, IF_tail, IL11, IMCp, INCENP_N, IRF-2BP1_2, ISG65-75, IclR, IncA, Inhibitor_I36, Inhibitor_I42, Innate_immun, Integrin_alpha2, Ion_trans_N, JHBP, Jnk-SapK_ap_N, KSHV_K1, KWG, Kinesin-relat_1, KorB, LEA_4, LEH, LRR_2, LRR_3, Lipoprotein_18, LppC, LysR_substrate, LytR_cpsA_psr, LytTR, MADF_DNA_bdg, MCE, MIG-14_Wnt-bd, MLTD_N, MORN_2, MOZART2, MRP, MVIN, MacB_PCD, MarR_2, Matrilin_ccoil, Med25, MerR-DNA-bind, Methyltransf_FA, Mga, MgtE_N, Milton, Mitoc_L55, Mlf1IP, Mu-like_gpT, NARG2_C, NAcGluc_Transf, NC, NERD, NESP55, NID, NIT, NOPS, NPIP, Nebulin, Neuralized, NfI_DNAbd_pre-N, Nrf1_DNA-bind, Nuc-transf, OCIA, Occludin_ELL, OppC_N, Orthoreo_P10, OstA, PBCV_basic_adap, PBP_like, PDDEXK_2, PEARLI-4, PGPGW, PIP49_C, POTRA_2, PP1c_bdg, PPK2, PRC, PaaA_PaaC, Pellino, Pentapeptide_2, Peptidase_A17, Peptidase_C11, Peptidase_M11, Peptidase_M23, Peptidase_S46, Peripla_BP_1, Phage_30_3, Phage_XkdX, Phage_fiber_2, Phage_head_chap, Phage_rep_org_N, Phasin_2, PhoU, Phospho_p8, Plant_NMP1, Plasmodium_HRP, Pollen_allerg_1, Potassium_chann, Pox_C4_C10, Prion_bPrPp, Prothymosin, Protocadherin, PspA_IM30, PurA, PyrI_C, RAP, RBB1NT, RDD, RGM_C, RHH_3, RNA_GG_bind, RNase_E_G, RNase_Zc3h12a, RPW8, RST,

	RWP-RK, Rb_C, RcbX, Receptor_2B4, Reg_prop, Ret_tiss, Root_cap, Rrf2, SAF, SARS_X4, SERTA, SH3_4, SK_channel, SLH, SNAP-25, SOUL, SR-25, STOP, SUFU, SUFU_C, SURF2, SapA, ScdA_N, Sec31, SecD_SecF, Self-incomp_S1, SerH, Serine_rich, Siah-Interact_N, Silic_transp, Sm_multidrug_ex, Smoothelin, SoxE, SoxG, SpoIIID, SpoOE-like, SpoVT_AbrB, Stork_head, SufE, Sulphotransf, Synapsin_C, TET_Cys_rich, TET_DSBH, TF_AP-2, TIL, TLV_coat, TPX2_importin, TatC, Terminase_6, Terminase_GpA, Thymopoietin, Thyroglobulin_1, Tissue_fac, Tmemb_55A, Tmemb_cc2, Toprim_N, Transposase_22, TylF, UPF0114, UPF0227, UPF0240, UPF0560, UPF0561, UPF0564, UPF0565, UnbV_ASPIC, UvrB, V-set, V-set_CD47, VASP_tetra, VHL, VWA_CoxE, VWA_N, VirC1, WXG100, Whirly, Wound_ind, Xylo_C, YHS, YajC, YceG, YceI, Ycf1, Ycf15, YlaC, YscO, YtxH, ZU5, ZapA, gpUL132, mTERF, plasmid_Toxin, rRNA_methylase, stn_TNFRSF12A, tify, ydhR, zf-LSD1, zf-RNPHF, zf-XS, zf-nanos

* Protein domains that are suggested to be metazoan-specific innovations by the Dollo parsimony are in red letters.

Supplementary Table S8. Numbers of *C. owczarzaki* Cys_2His_2 zinc fingers

No. ZFs in a protein	No. genes
1	21
2	3
3	9
4	4
5	1
7	3
8	1

Supplementary Table S9. Numbers of SH2 and PTB domains in *C. owczarzaki* and other eukaryotic lineages

Species	SH2	PTB
*T. thermophila**	1	0
N. gruberii	5	0
*D. discoideum**	14	0
R. oryzae	1	0
*S. cerevisiae**	1	0
C. owczarzaki	39	7
*M. brevicollis**	143	31
N. vectensis	29	27
*D. melanogaster**	34	10
*H. sapiens**	120	51

* data taken from[10]

Supplementary Notes

Supplementary Note 1: Genome structure

Analysis of synteny conservation

Conservation of gene order (conserved synteny) has been shown among many metazoan taxa, even between distantly related ones such as human and demosponge[1-4]. However, no clear conservation of synteny was detected between the genome of *C. owczarzaki* and that of *M. brevicollis*, *A. queenslandica* or *N. vectensis*.

Transposable elements

We screened the genome for the two classes of transposable element, these being the retrotransposons (both long terminal repeat (LTR) and non-LTR retrotransposons, which transpose via an RNA intermediate) and the canonical transposons (i.e. those that transpose only as DNA), following a previously published protocol[22]. The Repbase protein database[63] was used as the query library. Five LTR retrotransposon families (*C. owczarzaki chromovirus (Cocv) 1-5*) and four non-LTR retrotransposon families (*C. owczarzaki L1 (CoL) 1-4*) were identified. 14 families of canonical transposon were also identified (*C. owczarzaki bacterial transposon-like transposon* (*Cobalt*) *1-3*, *C. owczarzaki CACTA element* (*CoCACTA*) *1-2*, *C. owczarzaki MULE* (*Com*) *1-2*, *C. owczarzaki pogo-like element* (*Cop*) *1-5* and *C. owczarzaki Tc1-like element* (*CoTc*) *1-2*). Note that a family here refers to a group of nearly identical functional copies and non-functional trace sequences that are derived from a single transposable element, and that the families are further classified into superfamilies (*chromovirus, L1, Bacterial-like, CACTA, MULE, pogo, Tc1*) by their phylogenetic affinities. Genomic organizations of annotated transposable element families, their presence and absence in other eukaryotic genomes, and the characteristics of each family (e.g. length, copy number estimate and target site duplication pattern) are summarized in Supplementary Figures S1, S2,

and Table S1, respectively. Their sequences are deposited in the DNA Data Bank of Japan (DDBJ) under the accession numbers BR000974 to BR000996.

Orthologous families from all transposable element superfamilies have previously been identified in both fungal and metazoan genomes, consistent with their vertical inheritance within the Opisthokonta. The high-level nucleotide identity between the intra-element LTRs (Supplementary Table S1) indicates that all copies of full-length LTR retrotransposons are recent insertions in this genome. However one family, *Cocv4*, now appears to be non-functional, as the only full-length copy contains a 22bp deletion in its 3' LTR.

The approximate proportion of the genome composed from transposable element DNA was estimated by multiplying the copy number and the reported sequence length of each family. This approach gives a value of 1.48Mb, equating to 5.3% of the genome, for the 9 retrotransposon families. Of the 14 canonical transposon families, only four belonging to the *CACTA* and *MULE* superfamilies generate unique target site-duplications, allowing their copy numbers to be confidently determined. The other 10 transposon families produce 2bp, TA, duplications. As a result, it was only possible to provide upper and lower copy number estimates for those 10 transposon families, based upon the number of 5' and 3' termini observed (Supplementary Table S1). Accordingly, the proportion of the genome derived from transposon DNA is presented as a range. The lower and upper copy number estimates for each transposon family produce values of 1.00Mb and 1.43Mb respectively; these values equate to 3.7% and 5.2% of the genome. Combining the values for retrotransposons (5.3%) and canonical transposons (3.7-5.2%), an approximate total of 9.0-10.5% of the *C. owczarzaki* genome is of transposable element origin, which is much higher than that of the choanoflagellate *M. brevicollis* genome (~1%)[22].

Intergenic distance and evolution of gene regulation

The transition to multicellularity was probably accompanied by new or more complex mechanisms of regulation of gene expression. How regulatory regions shape genome architecture remains mostly unknown. However, it has been proposed that the distance between genes correlates, to some extent, with the amount of regulatory information contained in the non-protein-coding regions, especially in compact genomes such as *C. elegans* and *D. melanogaster*[64]. In these genomes, a positive correlation was observed between intergenic distances and the complexity of their expression patterns. Thus, we analyzed the intergenic distances for diverse eukaryotic taxa, examining whether there is any correlation between the functional category of a gene (to which regulatory complexity is linked to some extent) and the distance to its nearest upstream protein coding region (Supplementary Figure S3; detailed data in Figures S4 and S5). In *C. owczarzaki*, genes involved in receptor activity, transcriptional regulation and signaling processes have particularly large (33%, 66% and 39% higher geometric means, respectively) upstream intergenic regions compared to all the other genes. On the other hand, housekeeping genes such as ribosomal genes and metabolic genes have upstream intergenic regions with similar lengths to the others (less than 10% geometric mean difference). Although downstream regions follow a similar trend, the differences are not as large as in the upstream regions (Supplementary Figure S5). This pattern is seen across most of the taxa we analyzed, suggesting that the expression of transcription factors and molecules involved in communication and signaling is regulated by more complex transcriptional networks than those involved in housekeeping functions already in the early unikonts.

The mitochondrial genomes of filastereans

Animal mitochondrial DNAs (mtDNAs) are typically small (~16 kbp), circular-mapping molecules that encode 37 or fewer tightly packed genes (Supplementary Table S2). A notable exception is the placozoan *Trichoplax adhaerens*, which has mtDNAs with large intergenic regions and numerous introns[65]. Other unicellular relatives of the Metazoa have similarly non-animal like and structurally diverse mtDNAs, including the choanoflagellate *Monosiga brevicollis* (circular-mapping, with large, highly repetitive intergenic regions[66]), the ichthyosporean *Amoebidium parasiticum* (totaling >200 kbp and consisting of several hundred linear chromosomes with terminal-specific sequence patterns[66]), and *Ministeria vibrans* (Supplementary Figure S6; a linear genome with long inverted repeats). The mtDNA of C. *owczarzaki* makes no exception. With close to 200 kbp, it is the largest mtDNA in this comparison (Supplementary Figure S6). Similar to *M. brevicollis*, it contains highly repetitive intergenic regions, so the assembly remains incomplete. In Figure S6, the sequence is represented as a single linear contig, but the many repeats do not allow a confident prediction of the genome structure. Its gene content closely resembles those of *M. brevicollis* and *M. vibrans* (Supplementary Table S3). The large ORFs (Supplementary Figure S6B shown in green) found in the *C. owczarzaki* mtDNA are unrelated in sequence to those in mtDNA of other species and to mobile mitochondrial plasmids, thus their origin and potential function (if any) remains obscure. The only intron (group IB) in the *cox1* gene is atypical in including the gene for tRNA alanine.

Supplementary Note 2: Phylogenetic position of *C. owczarzaki*

Maximum likelihood (ML) and Bayesian Inference (BI) phylogenetic analyses were performed with several datasets (see Materials and Methods). In particular, two datasets were used: fMBH, which includes 141 proteins and 42,106 amino acid sites, and 145POP, which consists of 145 proteins and 37,146 amino acid sites. All four trees (fMBH-ML tree, fMBH-BI tree, 145POP-ML tree, and 145POP-BI tree) show *C. owczarzaki* as the sister lineage to choanoflagellates and metazoans with the maximum statistical supports, as previously shown[8,67] (Supplementary Figures S7 – S10).

Supplementary Note 3: Domain gain and loss in eukaryote evolution

The *C. owczarzaki* genome provides an unprecedented opportunity to better reconstruct the putative ancestral genome of the Holozoa. However, reconstructing an entire ancestral genome is challenging mainly due to mosaic proteins. A less complicated alternative is to infer the “domainomes", or the set of all protein domains, along a phylogenetic tree, by the use of Dollo parsimony[14,68].

We scanned the proteomes of 35 eukaryotes (14 metazoans, 1 choanoflagellate, *C. owczarzaki*, 9 fungi, the amoebozoan *D. discoideum* and 9 bikonts; see Materials and Methods for detail) for protein domains using the Pfam database. Dollo parsimony was then used to reconstruct ancestral domainomes. Results are summarized in Figure 2 of the main text. For a full list of domain gain and loss at the onset of Holozoa, Metazoa+Choanoflagellata, and Metazoa see Supplementary Table S4. In Tables S5 and S6, all GO term categories, except for very similar ones, that appear to have been enriched or depleted ($p < 0.001$) are shown together with the belonging Pfam domains.

The results confirmed the evolutionary trends described previously[14] such, as in metazoans, the emergence of apoptosis-related domains and a steady loss of domains constituting metabolic proteins. The data revealed two extensive increases of domains involved in gene regulation, one at the onset of Holozoa and the other at the onset of Metazoa (Supplementary Table S5). Some domains involved in signal transduction seem to have been significantly enriched at the onset of Holozoa. In contrast, transcription factors were significantly lost in both the choanoflagellate *M. brevicollis* and fungi (Supplementary Table S6), while many extracellular domains were lost in the lineages leading to *C. owczarzaki* and *M. brevicollis* (Supplementary Table S6).

It is worth mentioning that Dollo parsimony does not consider the possibility of lateral gene transfer, which may be affecting the reconstruction of domainnomes[14]. Therefore, in

Figure S11, a Venn diagram depicts the numbers of Pfam domains shared (or not shared) among four opisthokont lineages, metazoans, *M. brevicollis*, *C. owczarzaki* and fungi. Although 2299 domains are common among the four lineages, 903 domains appear to be metazoan-specific within the Opisthokonta (Supplementary Table S7 for the detail). The metazoan-specific domain repertoire inferred by the Dollo parsimony (Supplementary Table S4 Metazoa – gain; also shown by red letters in Table S7) and that by the Venn diagram (where the canonical parsimonious estimation is implied) (Supplementary Table S7) show a similar trend concerning the GO categories by which the domains are classified.

Supplementary Note 4: Protein domain enrichment analysis

Gene (or domain) duplication has been considered to be an important mechanism for the evolution of complex organisms[24,69]. Thus, protein domains that are enriched in metazoan genomes compared to those of non-metazoans may represent an important domain set involved in multicellularity and developmental processes of metazoans. We chose such domains from the Interpro database[25] by selecting domains that are statistically significantly enriched ($p<1.0e-20$ by Fisher's exact test; see Materials and Methods) in metazoan genomes compared to non-metazoans genomes excluding those of filastereans and choanoflagellates. Subsequently, whether these domains are also enriched or depleted in the genomes of *C. owczarzaki* and *M. brevicollis* was examined. The normalized numbers of genes containing such domains are shown by a heatmap (Supplementary Figure S12). In Figure S13, on which Figure 3 in the main text is based, redundant domains and domains present only in a single taxon are not exclusively shown, and domains were manually classified into 12 categories in terms of their known functions.

Supplementary Note 5: Analyses of Gene Families

Gene families of 12 selected biological categories are analyzed with a focus on the commonality and difference among gene repertoires of *C. owczarzaki*, *M. brevicollis*, and metazoans.

Cell adhesion

Cell adhesion systems are essential for metazoan multicellularity, mediating the physical contact and signal transduction between cells or cells and extracellular matrix (ECM)[12]. Of particular importance are the integrin-mediated adhesion and the cadherin-based machinery, which are involved in the focal adhesion and the spot adherens junction, respectively. The *C. owczarzaki* genome contains all the main components of the integrin-mediated adhesion machinery[16] (Supplementary Figure S14). The ECM components (e.g. fibronectins, laminins and collagens) are in contrast missing in this protist, although some related protein domains, such as laminin globular domain (Laminin G), and fibronectin type 2 and 3 domains, are present as modules composing other proteins. The only cadherin protein (CAOG_08574) of *C. owczarzaki* harbors 17 cadherin repeats, a signal peptide, a transmembrane segment and a short (83 amino acids) cytoplasmic domain, which has no clear sequence similarity to any known protein domain (Supplementary Figure S15). In the genome of the choanoflagellate *M. brevicollis*, on the other hand, integrin genes appear to have been secondarily lost while 23 cadherin genes are present[16,70].

C. owczarzaki also has some components of the dystrophin-associated glycoprotein complex (DGC), another cell-ECM adhesion system. In metazoan striated muscle tissue, the DGC connects the cytoskeleton with the ECM and transmits signals between them[71]. *C. owczarzaki* has the transmembrane receptor sarcoglycan (CAOG_04854) and the cytoplasmic components dystrophin (CAOG_03619) and syntrophin (CAOG_05815), while it lacks the

other receptor components (dystroglycans and sarcospan) and another cytoplasmic component (dystrobrevin). Homologs of other cell-cell adhesion molecules such as immunoglobulin-like cell adhesion molecule (IgCAM) and C-type lectins were not identified in the *C. owczarzaki* genome.

Transcription factors

The genome of *C. owczarzaki* contains a rich repertoire of transcription factors (TFs) (Supplementary Figure S16), including some TFs that play important roles in metazoan multicellularity, such as the T-box gene Brachyury (involved in mesoderm specification and blastopore determination), NF-κB (immune reaction), and Runx (control of cell differentiation and proliferation), which were previously thought to be metazoan innovations. Some of them thus clearly predate the origin of metazoa and are likely to have been recruited for other functions at the transition from unicellular to multicellular systems[16].

In Figure S16, the distribution of TFs among eumetazoan taxa is summarized. A general overview for each family is described previously[16], or described as follows.

COE

The transcription factor collier/olfactory-1/early B cell factors (COE) comprise the group F basic helix-loop-helix (bHLH) family. It has a single helix-loop-helix structure but lacks the basic amino acid region for DNA binding. Instead it has an N-terminal DNA-binding domain, which is conserved amongst all of the metazoan orthologs[72]. COE plays important roles in many aspects of metazoan development, including regulation of olfactory system, immune cell fates, and segmentation[73,74]. *C. owczarzaki* possesses a putative COE ortholog (CAOG_06470) although it lacks the HLH structure, whereas no homologs of this gene are found in the choanoflagellate *M. brevicollis*.

TEAD transcription factor

The TEA domain (TEAD) transcription factors are characterized by the presence of a TEA domain. They are specific to opisthokonts, but play different roles in fungi and metazoans. In fungi, the TEAD transcription factor Tec1 receives inputs from both the target of rapamycin (TOR) and mitogen-associated protein kinase (MAPK) pathways, coordinating the physiological response to pheromones and nutrients[75]. It can regulate target genes either alone, or together with the second transcription factor Ste12, which is specific to fungi[76]. In metazoans, the TEAD transcription factor Scalloped is the main binding partner of Yes-associated protein (YAP)/Yorkie, the hub of the Hippo pathway that controls cell growth and proliferation[77]. *C. owczarzaki* has a single TEAD transcription factor, which can substitute the function of the homologous *Drosophila* protein in the Hippo pathway[19]. *C. owczarzaki* lacks Ste12, but has YAP as well as other members of the Hippo pathway[19].

Zinc finger TFs

Zinc finger family is one of the most numerous and divergent eukaryotic TF families (Supplementary Figure S16)[78,79]. The classical Cys_2His_2 zinc fingers have typically the Cys-X_{2-4}-Cys-X_{12}-His-X_{3-5}-His sequence motif[56]. Sequence similarities between the various orthologs are very low, probably due to the weak evolutionary constraints on the sequence that is not involved in coordinating zinc ions. Only a few of them, such as Snail, Sp-1 and Gli, which in metazoans play fundamental roles in embryonic development[80], are well conserved across metazoan phyla[56,81]. We performed a HMMER search in *C. owczarzaki* genome with one Pfam model (PF00096), and manually inspected the motifs. We identified 42 Cys_2His_2 zinc finger genes in the genome of *C. owczarzaki* that were classified according to the number of zinc fingers contained in the encoding proteins (Supplementary Table S8). The conserved

TGEKP motif, which is likely involved in forming the structure and binding to DNA[78], was found in the linker sequences of four *C. owczarzaki* Cys_2His_2 zinc fingers (CAOG_06198, CAOG_06541, CAOG_06953, and CAOG_07968). Five *C. owczarzaki* zinc finger proteins have, in addition to the zinc finger domains, other known protein domains (Supplementary Figure S17).

Another large zinc finger TF family is comprised of the GATA factors. They have the DNA-binding sequence motif (Cys-X_{2-4}-Cys-X_n-Cys-X_2-Cys-basic region), which binds to the DNA motif (A/T)GATA(A/G)[79,82]. In metazoans, they are essential for many developmental processes including endoderm specification and cardiogenesis[82]. In fungi, however, they play divergent roles that appear totally different from those of metazoans such as nitrogen metabolism and mating-type switching[83]. Most metazoan GATA zinc fingers have 17-residues in the loop region of the domain, and two domains are arrayed in tandem in each protein. In fungi there are also GATA zinc fingers with 18-residue loops and they usually contain a single domain[83]. A leucine in the seventh position of the loop, which is considered important for the DNA-binding specificity[84], is conserved among most of the metazoan GATA zinc fingers, whereas it is not usually the case for fungi. The set of GATA zinc finger proteins encoded by the *C. owczarzaki* genome represents a kind of intermediate state between that of fungi and metazoans. The *C. owczarzaki* genome contains nine GATA factor genes with variable loop length (Supplementary Figure S18). Only one of them (CAOG_02090) has leucine in the seventh position of the loop like those in metazoans. Only one (CAOG_07963) has two zinc fingers, which are however not adjacent to each other in the protein. The rest have only one zinc finger, like those in fungi. CAOG_03534 is a putative MTA1 (metastasis-associated 1) homolog with the characteristic bromo-adjacent homology (BAH) domain. MTA1 is involved in carcinogenesis and the progression of tumors in mammals[85].

The $Zn(II)_2Cys_6$ family with six cysteines bound to two zinc ions is present in many eukaryotic genomes but is absent in metazoans and *M. brevicollis*. Their functions in fungi include sugar metabolism, ergosterol biosynthesis and meiosis[79]. *C. owczarzaki* has 12 genes belonging to this family. All of them contain the conserved zinc-binding motif Cys-X_2-Cys-X_6-Cys-$X_{5\text{-}16}$-Cys-X_2-Cys-$X_{6\text{–}8}$-Cys, a proline residue between the third and fourth cysteine that provides flexibility to the loop region, and a basic amino acid cluster between the second and third cysteine[86].

CSL

The transcription factor CBF1/RBP-Jκ/suppressor of hairless/LAG-1 (CSL) is the downstream effector in the Notch signaling pathway. This pathway was thought to be metazoan specific because the *bona fide* Delta or Notch transmembrane proteins are absent in non-metazoan genomes. However, the genomes of *M. brevicollis* and fungi[9,87] encode members of this TF family. In *S. cerevisiae*, CSLs are involved in cell adhesion and division[88]. *C. owczarzaki* also has a CSL (CAOG_08463), with the LAG-1-like DNA binding domain (known also as the Rel-homology region) followed by the β-trefoil domain and the weakly conserved IPT (immunoglobin-like fold shared by plexins and transcription factors) domain. The presence of CSL in *C. owczarzaki* is particularly interesting because its genome encodes some putatively primordial Notch signaling components (see Supplementary Note 5, Notch signaling).

RFX transcription factors

Regulatory factor X (RFX) transcription factors plays a crucial role in controlling flagellar development in the Metazoa[89]. The choanoflagellate *M. brevicollis*, which has one flagellum, possesses three RFX genes. *C. owczarzaki*, which lacks cilia, also has one RFX

gene (CAOG_00356), suggesting that the RFX transcription factors are not used for cilia development in *C. owczarzaki* (see also Supplementary Note 5, Flagellum), as it happens with some Fungi[90].

Heat Shock Factors

The heat shock response is mediated at the transcriptional level by cis-acting sequences called heat shock elements (HSE), which are present upstream of the heat shock protein (HSP) genes[91,92]. Heat shock factors (HSFs) can bind to the HSE and induce the expression of HSP genes. Trimerization of HSFs, which is important for its DNA-binding property, is mediated by arrays of hydrophobic heptad repeats A and B (HR-A and HR-B)[91]. Mammalian HSFs located on their sex chromosomes (HSFX and HSFY) exceptionally lack both HR-A and HR-B. *C. owczarzaki* has both types. One (CAOG_06646) has both HR-A and HR-B like the canonical mammalian HSFs (HSF1-4), whereas the second and third ones (CAOG_05916 and CAOG_00281) do not have either, like HSFX and HSFY.

Protein kinases

384 protein kinases (PKs) were found in the genome of *C. owczarzaki* (Supplementary Figure S22). They are classified into 8 groups, and further sub-classified into families, according to Hanks and Hunter[93] and http://www.kinase.com/. Half (197 out of 384) of *C. owczarzaki* PKs have putative metazoan orthologs, and 89 have putative *M. brevicollis* orthologs (Supplementary Figure S22). Out of the 197 PKs shared between *C. owczarzaki* and metazoans, 115 (33 if a highly expanded family with 82 members not included; see below) are missing in *M. brevicollis*.

Two families are particularly highly expanded in the filasterean lineage. First, the *C. owczarzaki* leucine rich repeat cytoplasmic serine/threonine kinase (CoLCSTK) family

contains 82 members, whose kinase domain sequences are closely related to each other. They are closely related to the IL1 receptor associated kinases (IRAKs) of metazoans. 68 of the 79 CoLSTK genes are made up of a putative serine/threonine kinase domain and 0-28 leucine rich repeats (LRRs), and often a C-terminal $Cys_3HisCys_4$ type zinc finger domain (Supplementary Figure S23). 11 of them show atypical architectures, suggesting domain duplication and shuffling during the diversification (Supplementary Figure S23). The other highly expanded family is the *C. owczarzaki* leucine rich receptor tyrosine kinase (CoLRYK) family, which contains 98 members. In contrast to the CoLCSTK genes, they are mostly receptor tyrosine kinases with numerous (4-46) extracellular LRRs, but this family contains genes with more divergent architectures, which seem to have been recently generated by frequent domain duplication, shuffling, and gene conversion[20].

The mitogen-activated protein kinase (MAPK) signaling cascade relays, within the cells, extracellular stimuli into nuclei and regulates various cellular responses[58,94,95]. Genes involved in this cascade are classified into three groups, MAPK, MAPK kinase (MAPKK), and MAPKK kinase (MAPKKK), depending principally on the order of regulation in the cascade. The *C. owczarzaki* genome contains 15 of the 21 cascade components that are in common between human and early-branching metazoans (either *T. adhaerens* or *A. queenslandica*), showing a greater conservation than in the choanoflagellate *M. brevicollis* (only 10 are shared) (Supplementary Figure S24). In contrast, the repertoire of *C. owczarzaki* receptor tyrosine kinases, which receive the extracellular signals and start the MAPK cascade within cells[96], is quite distinct from those of metazoans and choanoflagellates (see below).

Protein tyrosine kinases (TKs / PTKs) are, in metazoans, involved in cell-cell communication, cell differentiation and proliferation by transducing phospho-tyrosine signals initiated by extracellular ligands received by receptor molecules[97,98], and therefore play crucial roles in the multicellular development of metazoans[99]. Excluding two TKs

(CAOG_08142 and CAOG_08238) with incomplete kinase domains and one TK (CAOG_08366) on the supercontig 63 that may be a part of supercontig 16 with polymorphisms, we found 103 putative TKs in the *C. owczarzaki* genome. Ninety-two of them are predicted to be receptor TKs (RTKs) and 11 are classified as cytoplasmic TKs (CTKs)[20]. All *C. owczarzaki* TKs that are homologous to metazoan TKs are of cytoplasmic-type (homologs of Src, Csk, Abl, Fak, and Tec), whereas *C. owczarzaki* RTKs, which are not mapped to any metazoan RTKs, seem to have diverged specifically in the filasterean lineage[20].

Src homology 2 (SH2) domain and phosphotyrosine-binding (PTB) domain are involved in TK signal transduction by binding to a phosphorylated tyrosine residue[100]. It has been proposed that these domains highly expanded in number concurrently with the metazoan-type TK expansion during holozoan evolution[10]. The numbers of SH2 and PTB domains in the *C. owczarzaki* genome are consistent with this notion (Supplementary Table S9). However, the *C. owczarzaki* data indicate that the SH2 domain was extensively duplicated specifically in *M. brevicollis*. The genome also demonstrates that the PTB domains appeared before the divergence of filastereans, choanoflagellates and metazoans.

Protein tyrosine phosphatases

Seven classical protein tyrosine phosphatases (PTPs) ([101,102] for review) are present in the genome of *C. owczarzaki* (Supplementary Figure S25). One (CoPTP1) out of seven is classified as receptor-like PTPs (RPTPs), and six (CoPTP2-7) are classified as non-receptor PTPs (NRPTPs). Although the number of *C. owczarzaki* PTPs is smaller than that of *M. brevicollis*, which has 39 PTPs[10], their homologies to the metazoan PTPs are much greater. Six out of the seven of *C. owczarzaki* PTPs have putative human homologs (Supplementary Figure S25), whereas only four of the 39 *M. brevicollis* PTPs were found to be homologous to human genes[10]. Although frequent duplication of the PTP catalytic domain often obscures the

orthology between vertebrates and early-branching metazoans or protists[10,103], the domain architectures of *C. owczarzaki* PTPs are nearly identical to those of the putative human homologs (Supplementary Figure S25).

Notch signaling

Notch signaling is involved in cell-cell communication, cell differentiation, apoptosis and proliferation, which are all important features of metazoans[104-106]. Previous studies suggested Notch signaling to be metazoan-specific, mainly because of the lack of the *bona fide* ligand and receptor-encoding genes in non-metazoan genomes[9,57]. However, *M. brevicollis* possesses a gene encoding a receptor protein comprising, similarly to the metazoan Notch proteins, two Notch/Lin-12 repeats in its extracellular region and several ankyrin repeats in its cytoplasmic region[57]. *C. owczarzaki* has also two genes (CAOG_00333 and CAOG_06027) encoding proteins with similar domain architectures to this choanoflagellate protein (Supplementary Figure S19). Moreover, *C. owczarzaki* has a group of genes encoding proteins with a domain combination specific to Notch ligands: EGF-like repeats and Delta/Serrate/lag-2 (DSL) domains that are important for the interaction between the receptor and ligand of the Notch signaling[104,107]. Interestingly, some of these Notch ligand-like proteins are receptor tyrosine kinases, having the catalytic domains in their cytoplasmic region (Supplementary Figure S20). The absence of DSL domain in the *M. brevicollis* genome suggests that the origin of DSL domain antedates the divergence of filastereans and metazoans+choanoflagellates, and that it was lost in *M. brevicollis*. The interaction between Notch and Delta in metazoans is mediated by the DSL domain of delta and some specific EGF repeats of Notch[104], the latter lacking in the Notch-like proteins of *C. owczarzaki* and *M. brevicollis*. Therefore, a direct interaction between Notch-like protein and Delta-like protein in *C. owczarzaki* is unlikely. However, the possibility of an interaction between the Delta-like

protein and other proteins containing EGF domains, which are abundant in the genome of *C. owczarzaki* cannot be excluded.

Apart from receptors and ligands, many genes involved in this pathway are of ancient origin, while others seem to be metazoan innovations (Supplementary Figure S21). The most reasonable explanation would be that a co-option of already-existing genes and a generation of some new genes by domain shuffling have driven the assembly of a preliminary Notch signaling in the last common ancestor of metazoans[57].

Seven transmembrane receptors and G protein signaling

C. owczarzaki has a rich repertoire of seven-transmembrane (7TM) receptors, which are also referred to as G protein-coupled receptors (GPCRs) because most of them transmit signals by activating heterotrimeric G proteins[108]. They are involved in diverse signaling pathways that are important for cell communication and signaling[108]. 7TM receptors can be classified into eight families by their overall sequence homology and domain architecture[109]. *C. owczarzaki* has six out of these eight families, although it lacks the rhodopsin family like *M. brevicollis* (Supplementary Figure S26). The *C. owczarzaki* genome contains two families that are absent in *M. brevicollis*: the intimal thickness-related receptor (ITR) -like family and the Ocular albinism type 1 (OA1) -like family.

The origin of the three subunits (G_α, G_β, and G_γ) of heterotrimeric G proteins dates back to the origin of eukaryotes (Supplementary Figure S26). However, the orthologies of fungal G protein families to those of metazoans are unclear. The diversification of metazoan α subunit families, which are involved in coordinating the signals from divergent receptors to specific effectors[110,111] seems to be independent of the fungal Gα diversification[112]. The *C. owczarzaki* genome contains eight G_α genes, five of which are likely orthologs to metazoan G_α genes ($G_{\alpha s}$, CAOG_04667; $G_{\alpha v}$, CAOG_05446; $G_{\alpha i/t/o}$, CAOG_03262; $G_{\alpha 12/13}$, CAOG_01139; $G_{\alpha q}$,

CAOG_03395) (Supplementary Figure S26). This indicates that the divergence between the distinct metazoan Gα genes antedates the split of filastereans and metazoans+choanoflagellates. Similarly the G_{β}-1-4 and G_{β}-5 groups also diverged before this split.

G protein signaling is regulated also by a group of proteins sharing a domain called regulator of G protein signaling (RGS). RGS proteins modulate the G protein activity by interacting with G_{α}[113]. We classified them into 13 families according to the protein domain architectures and sequence similarity between the RGS domains (Supplementary Figures S26 and S27). The *C. owczarzaki* genome shows a stronger conservation of these families (seven out of 13) than the *M. brevicollis* genome, which retains only three.

G protein signaling is also mediated by second messengers such as cyclic nucleotides. Cyclic nucleotide phosphodiesterases (PDEs) regulate the spatiotemporal concentration of cyclic nucleotides, and thus they are important for interpreting extracellular signals to specific intracellular signals[114]. *C. owczarzaki* has three PDEs, two of which belong to the PDE2 family (CAOG_04883) and the PDE7 family (CAOG_01392), respectively.

Other signal transduction systems

The *C. owczarzaki* genome does not encode proteins involved in the TGF-β signaling, i.e. its ligand, receptor, the downstream transcription factors Smad, Jun, and Fos, the ligand inhibitor Noggin, and the other downstream target c-Jun N-terminal kinase (JNK). It is likely that this pathway is a truly metazoan innovation[9,115].

Main players of the canonical Hedgehog (Hh) signaling are lacking in the *C. owczarzaki* genome. *C. owczarzaki* does not have the ligand Hh, Dispatched (involved in ligand secretion), or Patched and Smoothened (involved in ligand reception). The Hh-like protein identified in *M. brevicollis* is split up into two different proteins: one having a sequence similarity with the

C-terminal part (Hint or Hog domain) of Hh, and another possessing the N-terminal part (Hedge domain), which appears as a large transmembrane protein[9,116]. However, neither Hint nor Hedge domain is encoded by the *C. owczarzaki* genome. One of the main targets of the Hh signaling is the transcription factor Gli[117]. Although *C. owczarzaki* has a protein (CAOG_06541) containing five Cys_2His_2 zinc fingers that are closely related to those of Gli, whether it is the *bona* fide Gli ortholog remains unclear.

Similarly, no receptor or ligand genes were found for Wnt signaling in the genome of *C. owczarzaki*. None of the 7TM receptors of *C. owczarzaki* is closely-related to the Frizzled family (see Supplementary Figure S26), which is the receptor of canonical Wnt signaling. Moreover, none of its 7TM receptors have the specific Fz domain, which is involved in the interaction with the ligand Wnt. Because the amoebozoan *D. discoideum* has a putative Frizzled homolog with the Fz domain[118], the receptor Frizzled may have been secondary lost in both *C. owczarzaki* and *M. brevicollis*.

The JAK/STAT signaling is involved in relaying various cellular signals of metazoans such as extracellular growth factors. *C. owczarzaki* and *M. brevicollis* have a STAT transcription factor (see Supplementary Figure S16). However, both of them lack an ortholog of the cytoplasmic tyrosine kinase (CTK) JAK, even though they have the orthologs of most metazoan CTKs[10,20,119]. *M. brevicollis* has a possible JAK homolog with a domain architecture shared with metazoan JAKs; yet the orthology of its kinase domain to those of metazoans is not supported by phylogenetic analyses[20].

The Hippo signaling is involved in controlling organ size in metazoans, coordinating cell proliferation and apoptosis. The main players of this pathway, i.e., the kinases Hippo and Warts, the co-activator Yorkie, and the transcription factor Scalloped, are present in the *C. owczarzaki* genome[19]. Moreover, their functions and biochemical properties are also conserved between *C. owczarzaki* and *D. melanogaster*[19]. It has been recently suggested that

the GPCR pathway, in which the signal transducer G proteins are well-conserved between metazoans and *C. owczarzaki* (see Supplementary Note 5, Seven transmembrane receptors and G protein signaling), is involved in the Hippo signal transduction[120].

The Akt pathway is an ancient eukaryotic signaling system broadly found across eukaryotes (Supplementary Figure S36). In metazoans, it is particularly important in controlling cell proliferation and cell growth[121] and there are a number of animal-specific substrates that Akt regulates[121]. The repertoire of *C. owczarzaki* genes involved in Akt signaling is well conserved like other eukaryotes including non-holozoans (Supplementary Figure S36). However, several genes such as the E3 ubiquitin-protein ligase CBL, Growth factor receptor-bound protein 2 (GRB2), and Son of sevenless homolog, appear to be holozoan-specific.

Meiotic genes

There is no clear evidence for, or against, meiosis and sex in *C. owczarzaki*. We therefore searched the *C. owczarzaki* genome for a set of 29 core meiotic genes[122-125], some of which are also involved in DNA repair and mitosis. The genome of *C. owczarzaki* contains 27 (or 28 if a putative Rec8 ortholog is considered; see below) out of these 29 genes, a similar repertoire to those of metazoans, choanoflagellates, and other eukaryotes (Supplementary Figure S28). Included among them are *spo11-1* (CAOG_05659) and *spo11-3* (CAOG_03056), which catalyze DNA double-strand breaks specifically during the early stage of meiosis in metazoans, fungi, and plants[125,126]. Both *spo11-1* and *spo11-3* are also present in the *M. brevicollis* genome, while *spo11*-3 is absent in metazoans. *C. owczarzaki* also has a possible ortholog of the Rec8 gene (CAOG_00985), which encodes a protein involved in meiosis-specific chromatid cohesion, although it seems to have been lost in several basal metazoan lineages and *M. brevicollis*.

The foregoing analyses strongly suggest that *C. owczarzaki* undergoes meiosis. It is worth mentioning that *C. owczarzaki* has a homolog of HAPLESS2-Generative cell specific 1 (HAP2-GCS1) (CAOG_07409), which appears to be crucial for fusing the gamete plasma membranes in pre-opisthokonts[127]. *C. owczarzaki* may occasionally require amphimixis in order to purge deleterious mutations introduced by retrotransposons, as proposed for choanoflagellates[123] (see also Supplementary Note 1, Transposable elements).

Cell cycle regulators

Control of cell proliferation is critical to the survival of single-cellular organisms. Cyclins play a major role in cell cycle progression across all eukaryotes. In general, a cyclin forms a complex with a cyclin-dependent kinase (CDK) and regulates its activity, oscillating his own abundance during cell cycle[128]. *C. owczarzaki* has three orthologs (cyclin A, B, and E) of the four cyclin classes (cyclin A, B, D, and E) that are crucially involved in cell cycle regulation in bilaterians (Supplementary Figure S29). In contrast, *M. brevicollis* has secondarily lost both cyclin D and cyclin E, which are considered to be the major regulators of the G1/S transition. *C. owczarzaki* also has a nearly complete (9 out of 11 analyzed) repertoire of CDKs and other kinases involved in cell cycle regulation. Transcription factors involved in cell cycle regulation are also well conserved in the *C. owczarzaki* genome, although some of them (e.g. the E2F family) have specifically diversified in metazoans[129]. In particular, the Myc/Max/Mad network, which is involved in the transcriptional control of cell behavior such as cell proliferation and apoptosis[130], seems to have emerged at the onset of the Holozoa[18].

Flagellum

All extant eukaryotes may have or once had a flagellum or cilium[131]. A conserved gene complement for flagellum has been described among distantly-related eukaryotes[132-135]. We

examined whether *C. owczarzaki*, which lacks a flagellum in its known life cycle, retains or has secondarily lost the gene complement for flagellum. We additionally surveyed the genomes of four eukaryotes (*Chlamydomonas reinhardtii*, *Naegleria gruberi*, *Monosiga brevicollis*, and *Homo sapiens*), which have different types of flagellated cells, and compared their flagellum gene complements with that of *C. owczarzaki*. We found that more than 75% of the 117 genes involved in flagellum construction and motility[133-135] were lost in the *C. owczarzaki* genome (Supplementary Figure S30). Critical losses include δ- and ε- tubulins and the set of genes involved in intraflagellar transport, basal body construction, and tubulin-tyrosine ligation. Moreover, *C. owczarzaki* has lost some motor protein kinesins, specifically those (kinesin 2, 9, and 13) whose flagellar functions were suggested[26]. Kinesin 17, which is *in silico* predicted to have a flagellar role[136], is also lacking in *C. owczarzaki* genome. On the other hand, *C. owczarzaki* has a Regulatory factor X (RFX) transcription factor, the major transcriptional regulator of flagellar genes in certain metazoans[137,138] (see Supplementary Note 5, Transcription factors). This indicates that RFX already existed in holozoan ancestors and was recruited much later for flagellar gene regulation in metazoans[138].

RNA-binding protein genes

RNA-binding proteins (RBPs) regulate all aspects of post-transcriptional RNA biogenesis and have key roles in regulation of gene expression, a fundamental process during development in multicellular organisms[139]. RBPs also control the biogenesis and function of non-coding RNAs (ncRNAs), such as micro-RNAs (miRNAs), which can influence gene expression at both transcriptional and post-transcriptional levels, notably during development and in stem cells[140]. Many RBPs share common protein domains that are bound to RNA[141,142], including the RNA recognition motif (RRM), the heterogeneous nuclear RNP K-homology domain (KH) domain, the DEAD/DExH-box helicase domain (DEAD-box), and the double-

stranded RNA-binding domain (DsRM). The genome of *C. owczarzaki* encodes many proteins that contain one or more of these domains, as well as those that have sequence similarities to other known RBPs containing other types of domains, such as PUF, zinc fingers, Sm, Piwi, and PAZ domains[62,141]. We identified 184 putative RBPs in *C. owczarzaki*: 68 RRM, 12 KH, 62 DEAD-box, 2 dsRM, and 40 proteins with other RNA-binding domains (Supplementary Figure S31).

The data suggest that many RBP families had already evolved before the emergence of unikonts, and then another series of gene duplications occurred during metazoan evolution, most likely before the separation of sponge and eumetazoans (Supplementary Figure S32). It is also noteworthy that several genes involved in synthesis and functioning of non-coding RNA (ncRNA) in eukaryotes, e.g. micro-RNA (miRNA) and Piwi-interacting RNA (piRNA), seem to have been lost in *C. owczarzaki* and *M. brevicollis* (Supplementary Figure S32).

Neuronal genes

Neurosecretion

Peptidylglycine α-amidating monooxygenase (PAM) is found in *C. owczarzaki,* metazoans, and the green algae *C. reinhardtii* and *V. carteri*, but absent in *M. brevicollis* (Supplementary Figure S33). In metazoans this protein is multifunctional with two catalytic domains: a peptidylglycine α-hydroxylating monooxygenase and a peptidyl-α-hydroxyglycine α-amidatinglyase, which sequentially catalyzes the conversion of peptides into active α-amidated products for secretion[143]. The *C. owczarzaki* PAM homolog appears to be functional, possessing both domains typically found in metazoans. We were able to identify only one protein convertase in *C. owczarzaki*: Subtilisin/Kexin type 1 protein (PC1/3), which is a key component in the processing of active neuropeptides and hormones. We also identified a Cathepsin L gene, a lysosomal endopeptidase that is present throughout eukaryotes, but is

absent in fungi. Similarly, *C. owczarzaki* has a carboxypeptidase-D homolog, which shows a similar phylogenetic distribution as Cathepsin L. Further, we found both leukotriene A4 hydrolase and glutaminyl-peptide cyclotransferase genes in *C. owczarzaki*, which are also broadly distributed throughout eukaryotes.

Presynaptic proteins

C. owczarzaki has a wide variety of genes that encode presynaptic proteins in metazoans (Supplementary Figure S34). The proteins involved in cell-cell adhesion and cell-extracellular matrix interactions are mostly holozoan specific. *C. owczarzaki* has one homolog of Cadherin, which is involved in cell adhesion (see Supplementary Note 5, Cell adhesion). *C. owczarzaki* also has a homolog of Cortactin, an actin binding protein that promotes actin cytoskeletal rearrangements and polymerization when activated by an external stimulus. Cortactin appears to be a holozoan specific innovation. Homologs of Cytohesin appear to have been lost in fungi. We also identified a protein tyrosine phosphatase receptor type F and a Neurexin type I/II/III protein, although without the calcium-binding receptor EGF-like domain (SMART accession number SM00181).

The proteins involved in synaptic vesicles and the endocytosis of synaptic vesicles in metazoans generally show a broad distribution across eukaryotes and thus likely represent co-option of these proteins for these specific tasks in the neuronal system. *C. owczarzaki* has homologs of all these proteins except synaptic vesicle glycoprotein 2 (SV2), transmembrane protein 163, and Synapsin.

There are homologs of the presynaptic signaling proteins, Abl tyrosine kinase, Ca^{2+} and integrin binding, and GPCR-kinase interactor in the *C. owczarzaki* genome. All of these appear to be holozoan innovations and are also present in the *M. brevicollis* genome.

Homologs of the presynaptic small GTPase proteins are broadly distributed across eukaryotes. There are only a few proteins that appear to be holozoan specific, such as Ras p21 activator 1, Rap Guanine nucleotide exchange factor 4, and Rab 3A interacting protein. However, the Rab 3A interacting protein is missing in *C. owczarzaki*.

The SNARE (soluble NSF attachment protein receptor) proteins are also broadly distributed. Of the proteins that are associated with synapses, only Syntaxin binding 4 protein is holozoan specific, although it is not present in *C. owczarzaki*,

Some of the presynaptic trafficking regulatory proteins, such as BAI1-associated 3, Synaptogyrin, and Synaptoporin/Synaptophysin, appear to be holozoan innovations. The synaptosomal-associated protein (SNAP)-associated is found in unikonts but appears to have been lost in fungi.

Postsynaptic proteins

C. owczarzaki has many genes that encode postsynaptic proteins. These protein groups can be divided into three functional groups (Supplementary Figure S35). Most of the postsynaptic proteins analyzed here are holozoan or metazoan specific. Only a few can be found in non-holozoans and in these cases they are genes that appear to have been co-opted for postsynaptic use.

The Discs large (DLG) of *C. owczarzaki* is similar in domain structure to the metazoans, but lacks the receptor targeting domain L27 and the N-terminal polyubiquitination domains, which are usually found in the metazoan DLGs[144]. Consistent with this observation, the DLG-associated 1 protein was also identified in *C. owczarzaki*. We also found a SH3 and Multiple Ankyrin Repeats (SHANK), a Scribbled homolog, a protein interacting with the protein kinase C alpha (PRKCA), and a Homer homolog in the *C. owczarzaki* genome. The Homer homolog of *C. owczarzaki* has a similar domain structure to those of metazoans, except for the

Med15 domain, which is only present in the *C. owczarzaki* protein. In the *C. owczarzaki* genome, there are several homologs of proteins that are involved in postsynaptic signaling, such as the signal-induced proliferation-associated 1-like and cysteine-rich PDZ-binding, both of which appear to be holozoan specific. Therefore, the origin of post-synaptic components may antedate the split of filastereans from choanoflagellates and metazoans[145].

Supplementary References

53. Levesque, C.A. *et al.* Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire. *Genome Biol* **11**, R73 (2010).

54. Dickinson, D.J., Nelson, W.J. & Weis, W.I. A polarized epithelium organized by beta- and alpha-catenin predates cadherin and metazoan origins. *Science* **331**, 1336-9 (2011).

55. Grimson, M.J. *et al.* Adherens junctions and beta-catenin-mediated cell signalling in a non-metazoan organism. *Nature* **408**, 727-31 (2000).

56. Shimeld, S.M. C2H2 zinc finger genes of the Gli, Zic, KLF, SP, Wilms' tumour, Huckebein, Snail, Ovo, Spalt, Odd, Blimp-1, Fez and related gene families from Branchiostoma floridae. *Dev Genes Evol* **218**, 639-49 (2008).

57. Gazave, E. *et al.* Origin and evolution of the Notch signalling pathway: an overview from eukaryotic genomes. *BMC Evol Biol* **9**, 249 (2009).

58. Dhanasekaran, D.N. & Johnson, G.L. MAPKs: function, regulation, role in cancer and therapeutic targeting. *Oncogene* **26**, 3097-9 (2007).

59. Gustin, M.C., Albertyn, J., Alexander, M. & Davenport, K. MAP kinase pathways in the yeast Saccharomyces cerevisiae. *Microbiol Mol Biol Rev* **62**, 1264-300 (1998).

60. Wilkinson, M.G. & Millar, J.B. Control of the eukaryotic cell cycle by MAP kinase signaling pathways. *Faseb J* **14**, 2147-57 (2000).

61. Oka, Y., Saraiva, L.R., Kwan, Y.Y. & Korsching, S.I. The fifth class of Galpha proteins. *Proc Natl Acad Sci U S A* **106**, 1484-9 (2009).

62. Kerner, P., Degnan, S.M., Marchand, L., Degnan, B.M. & Vervoort, M. Evolution of RNA-binding proteins in animals: insights from genome-wide analysis in the sponge Amphimedon queenslandica. *Mol Biol Evol* **28**, 2289-303 (2011).

63. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-7 (2005).

64. Nelson, C.E., Hersh, B.M. & Carroll, S.B. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol* **5**, R25 (2004).

65. Dellaporta, S.L. *et al.* Mitochondrial genome of Trichoplax adhaerens supports placozoa as the basal lower metazoan phylum. *Proc Natl Acad Sci U S A* **103**, 8751-6 (2006).

66. Burger, G., Forget, L., Zhu, Y., Gray, M.W. & Lang, B.F. Unique mitochondrial genome architecture in unicellular relatives of animals. *Proc Natl Acad Sci U S A* **100**, 892-7 (2003).

67. Ruiz-Trillo, I. *et al.* *Capsaspora owczarzaki* is an independent opisthokont lineage. *Curr. Biol.* **14**, R946-7 (2004).

68. Farris, J.S. Phylogenetic Analysis Under Dollo's Law. *Syst. Zool.* **26**, 77-88 (1977).

69. Zhang, J. Evolution by gene duplication: an update. *Trends Ecol Evol* **18**, 292-298 (2003).

70. Abedin, M. & King, N. The premetazoan ancestry of cadherins. *Science* **319**, 946-8 (2008).

71. Lapidos, K.A., Kakkar, R. & McNally, E.M. The dystrophin glycoprotein complex: signaling strength and integrity for the sarcolemma. *Circ Res* **94**, 1023-31 (2004).

72. Jackson, D.J. *et al.* Developmental expression of COE across the Metazoa supports a conserved role in neuronal cell-type specification and mesodermal development. *Dev Genes Evol* **220**, 221-34 (2010).

73. Crozatier, M., Valle, D., Dubois, L., Ibnsouda, S. & Vincent, A. Collier, a novel regulator of Drosophila head development, is expressed in a single mitotic domain. *Curr Biol* **6**, 707-18 (1996).

74. Wang, M.M. & Reed, R.R. Molecular cloning of the olfactory neuronal transcription factor Olf-1 by genetic selection in yeast. *Nature* **364**, 121-6 (1993).

75. Brückner, S. *et al.* The TEA transcription factor Tec1 links TOR and MAPK pathways to coordinate yeast development. *Genetics* **189**, 479-94 (2011).

76. Heise, B. *et al.* The TEA transcription factor Tec1 confers promoter-specific gene regulation by Ste12-dependent and -independent mechanisms. *Eukaryot Cell* **9**, 514-31 (2010).

77. Zhang, L. *et al.* The TEAD/TEF family of transcription factor Scalloped mediates Hippo signaling in organ size control. *Dev Cell* **14**, 377-87 (2008).

78. Laity, J.H., Lee, B.M. & Wright, P.E. Zinc finger proteins: new insights into structural and functional diversity. *Curr Opin Struct Biol* **11**, 39-46 (2001).

79. MacPherson, S., Larochelle, M. & Turcotte, B. A fungal family of transcriptional regulators: the zinc cluster proteins. *Microbiol Mol Biol Rev* **70**, 583-604 (2006).

80. Zhao, C. & Meng, A. Sp1-like transcription factors are regulators of embryonic development in vertebrates. *Dev Growth Differ* **47**, 201-11 (2005).

81. Schaeper, N.D., Prpic, N.M. & Wimmer, E.A. A clustered set of three Sp-family genes is ancestral in the Metazoa: evidence from sequence analysis, protein domain structure, developmental expression patterns and chromosomal location. *BMC Evol Biol* **10**, 88 (2010).

82. Patient, R.K. & McGhee, J.D. The GATA family (vertebrates and invertebrates). *Curr Opin Genet Dev* **12**, 416-22 (2002).

83. Scazzocchio, C. The fungal GATA factors. *Curr Opin Microbiol* **3**, 126-31 (2000).

84. Ravagnani, A. *et al.* Subtle hydrophobic interactions between the seventh residue of the zinc finger loop and the first base of an HGATAR sequence determine promoter-specific recognition by the Aspergillus nidulans GATA factor AreA. *Embo J* **16**, 3974-86 (1997).

85. Toh, Y. & Nicolson, G.L. The role of the MTA family and their encoded proteins in human cancers: molecular functions and clinical implications. *Clin Exp Metastasis* **26**, 215-27 (2009).

86. Marmorstein, R., Carey, M., Ptashne, M. & Harrison, S.C. DNA recognition by GAL4: structure of a protein-DNA complex. *Nature* **356**, 408-14 (1992).

87. Převorovský, M., Půta, F. & Folk, P. Fungal CSL transcription factors. *BMC Genomics* **8**, 233 (2007).

88. Převorovský, M. *et al.* Cbf11 and Cbf12, the fission yeast CSL proteins, play opposing roles in cell adhesion and coordination of cell and nuclear division. *Exp Cell Res* **315**, 1533-47 (2009).

89. Thomas, J. *et al.* Transcriptional control of genes involved in ciliogenesis: a first step in making cilia. *Biol Cell* **102**, 499-513 (2010).

90. Chu, J.S., Baillie, D.L. & Chen, N. Convergent evolution of RFX transcription factors and ciliary genes predated the origin of metazoans. *BMC Evol Biol* **10**, 130 (2010).

91. Åkerfelt, M., Morimoto, R.I. & Sistonen, L. Heat shock factors: integrators of cell stress, development and lifespan. *Nat Rev Mol Cell Biol* **11**, 545-55 (2010).

92. Åkerfelt, M., Trouillet, D., Mezger, V. & Sistonen, L. Heat shock factors at a crossroad between stress and development. *Ann N Y Acad Sci* **1113**, 15-27 (2007).

93. Hanks, S.K. & Hunter, T. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.* **9**, 576-596 (1995).

94. Chen, Z. *et al.* MAP kinases. *Chem Rev* **101**, 2449-76 (2001).

95. Seger, R. & Krebs, E.G. The MAPK signaling cascade. *Faseb J* **9**, 726-35 (1995).

96. Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **103**, 211-25 (2000).

97. Fantl, W.J., Johnson, D.E. & Williams, L.T. Signalling by receptor tyrosine kinases. *Annu. Rev. Biochem.* **62**, 453-481 (1993).

98. van der Geer, P., Hunter, T. & Lindberg, R.A. Receptor protein-tyrosine kinases and their signal transduction pathways. *Annu. Rev. Cell Biol.* **10**, 251-337 (1994).

99. Gerhart, J. 1998 Warkany lecture: signaling pathways in development. *Teratology* **60**, 226-39 (1999).

100. Schlessinger, J. & Lemmon, M.A. SH2 and PTB domains in tyrosine kinase signaling. *Sci Signal* **2003**, RE12 (2003).

101. Alonso, A. *et al.* Protein tyrosine phosphatases in the human genome. *Cell* **117**, 699-711 (2004).

102. Tonks, N.K. Protein tyrosine phosphatases: from genes, to function, to disease. *Nat. Rev. Mol. Cell Biol.* **7**, 833-46 (2006).

103. Ono, K., Suga, H., Iwabe, N., Kuma, K. & Miyata, T. Multiple Protein Tyrosine Phosphatases in Sponges and Explosive Gene Duplication in the Early Evolution of Animals Before the Parazoan- Eumetazoan Split. *J. Mol. Evol.* **48**, 654-662 (1999).

104. Artavanis-Tsakonas, S., Rand, M.D. & Lake, R.J. Notch signaling: cell fate control and signal integration in development. *Science* **284**, 770-6 (1999).

105. Mumm, J.S. & Kopan, R. Notch signaling: from the outside in. *Dev Biol* **228**, 151-65 (2000).

106. Lai, E.C. Notch signaling: control of cell communication and cell fate. *Development* **131**, 965-73 (2004).

107. Cordle, J. *et al.* A conserved face of the Jagged/Serrate DSL domain is involved in Notch trans-activation and cis-inhibition. *Nat Struct Mol Biol* **15**, 849-57 (2008).

108. Pierce, K.L., Premont, R.T. & Lefkowitz, R.J. Seven-transmembrane receptors. *Nat Rev Mol Cell Biol* **3**, 639-50 (2002).

109. Nordström, K.J., Sallman Almén, M., Edstam, M.M., Fredriksson, R. & Schiöth, H.B. Independent HHsearch, Needleman--Wunsch-based, and motif analyses reveal the overall hierarchy for most of the G protein-coupled receptor families. *Mol Biol Evol* **28**, 2471-80 (2011).

110. Birnbaumer, L. Receptor-to-effector signaling through G proteins: roles for $\beta\gamma$ dimers as well as α subunits. *Cell* **71**, 1069-1072 (1992).

111. Simon, M.I., Strathmann, M.P. & Gautam, N. Diversity of G proteins in signal transduction. *Science* **252**, 802-808 (1991).

112. Suga, H. *et al.* Extensive gene duplication in the early evolution of animals before the parazoan-eumetazoan split demonstrated by G proteins and protein tyrosine kinases from sponge and hydra. *J. Mol. Evol.* **48**, 646-653 (1999).

113. Willars, G.B. Mammalian RGS proteins: multifunctional regulators of cellular signalling. *Semin Cell Dev Biol* **17**, 363-76 (2006).

114. Conti, M. & Beavo, J. Biochemistry and physiology of cyclic nucleotide phosphodiesterases: essential components in cyclic nucleotide signaling. *Annu Rev Biochem* **76**, 481-511 (2007).

115. Adamska, M. *et al.* Wnt and TGF-beta expression in the sponge Amphimedon queenslandica and the origin of metazoan embryonic patterning. *PLoS One* **2**, e1031 (2007).

116. Snell, E.A. *et al.* An unusual choanoflagellate protein released by Hedgehog autocatalytic processing. *Proc Biol Sci* **273**, 401-7 (2006).

117. Dahmane, N., Lee, J., Robins, P., Heller, P. & Ruiz i Altaba, A. Activation of the transcription factor Gli1 and the Sonic hedgehog signalling pathway in skin tumours. *Nature* **389**, 876-81 (1997).

118. Eichinger, L. *et al.* The genome of the social amoeba Dictyostelium discoideum. *Nature* **435**, 43-57 (2005).

119. Suga, H. *et al.* Ancient divergence of animal protein tyrosine kinase genes demonstrated by a gene family tree including choanoflagellate genes. *FEBS Lett.* **582**, 815-8 (2008).

120. Yu, F.X. *et al.* Regulation of the Hippo-YAP pathway by G-protein-coupled receptor signaling. *Cell* **in press**(2012).

121. Manning, B.D. & Cantley, L.C. AKT/PKB signaling: navigating downstream. *Cell* **129**, 1261-74 (2007).

122. Schurko, A.M. & Logsdon, J.M., Jr. Using a meiosis detection toolkit to investigate ancient asexual "scandals" and the evolution of sex. *Bioessays* **30**, 579-89 (2008).

123. Carr, M., Leadbeater, B.S. & Baldauf, S.L. Conserved meiotic genes point to sex in the choanoflagellates. *J Eukaryot Microbiol* **57**, 56-62 (2010).

124. Malik, S.B., Pightling, A.W., Stefaniak, L.M., Schurko, A.M. & Logsdon, J.M., Jr. An expanded inventory of conserved meiotic genes provides evidence for sex in Trichomonas vaginalis. *PLoS One* **3**, e2879 (2008).

125. Malik, S.B., Ramesh, M.A., Hulstrand, A.M. & Logsdon, J.M., Jr. Protist homologs of the meiotic Spo11 gene and topoisomerase VI reveal an evolutionary history of gene duplication and lineage-specific loss. *Mol Biol Evol* **24**, 2827-41 (2007).

126. Keeney, S., Giroux, C.N. & Kleckner, N. Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* **88**, 375-84 (1997).

127. Wong, J.L. & Johnson, M.A. Is HAP2-GCS1 an ancestral gamete fusogen? *Trends Cell Biol* **20**, 134-41 (2009).

128. Murray, A.W. Recycling the cell cycle: cyclins revisited. *Cell* **116**, 221-34 (2004).

129. Cao, L. *et al.* The ancient function of RB-E2F pathway: insights from its evolutionary history. *Biol Direct* **5**, 55 (2010).

130. Grandori, C., Cowley, S.M., James, L.P. & Eisenman, R.N. The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu Rev Cell Dev Biol* **16**, 653-99 (2000).

131. Cavalier-Smith, T. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int J Syst Evol Microbiol* **52**, 297-354 (2002).

132. Carvalho-Santos, Z., Azimzadeh, J., Pereira-Leal, J.B. & Bettencourt-Dias, M. Evolution: Tracing the origins of centrioles, cilia, and flagella. *J Cell Biol* **194**, 165-75 (2011).

133. Pazour, G.J., Agrin, N., Leszyk, J. & Witman, G.B. Proteomic analysis of a eukaryotic cilium. *J Cell Biol* **170**, 103-13 (2005).

134. Fritz-Laylin, L.K. *et al.* The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell* **140**, 631-42 (2010).

135. Merchant, S.S. *et al.* The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**, 245-50 (2007).

136. Wickstead, B., Gull, K. & Richards, T.A. Patterns of kinesin evolution reveal a complex ancestral eukaryote with a multifunctional cytoskeleton. *BMC Evol Biol* **10**, 110 (2010).

137. Swoboda, P., Adler, H.T. & Thomas, J.H. The RFX-type transcription factor DAF-19 regulates sensory neuron cilium formation in C. elegans. *Mol Cell* **5**, 411-21 (2000).

138. Piasecki, B.P., Burghoorn, J. & Swoboda, P. Regulatory Factor X (RFX)-mediated transcriptional rewiring of ciliary genes in animals. *Proc Natl Acad Sci U S A* **107**, 12969-74 (2010).

139. Lasko, P. Gene regulation at the RNA layer: RNA binding proteins in intercellular signaling networks. *Sci STKE* **2003**, RE6 (2003).

140. Ghildiyal, M. & Zamore, P.D. Small silencing RNAs: an expanding universe. *Nat Rev Genet* **10**, 94-108 (2009).

141. Anantharaman, V., Koonin, E.V. & Aravind, L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res* **30**, 1427-64 (2002).

142. Lunde, B.M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol* **8**, 479-90 (2007).

143. Eipper, B.A., Milgram, S.L., Husten, E.J., Yun, H.Y. & Mains, R.E. Peptidylglycine alpha-amidating monooxygenase: a multifunctional protein with catalytic, processing, and routing domains. *Protein Sci* **2**, 489-97 (1993).

144. De Mendoza, A., Suga, H. & Ruiz-Trillo, I. Evolution of the MAGUK protein gene family in premetazoan lineages. *BMC Evol. Biol.* **10**, 93 (2010).

145. Alié, A. & Manuel, M. The backbone of the post-synaptic density originated in a unicellular ancestor of choanoflagellates and metazoans. *BMC Evol Biol* **10**, 34 (2010).